Computational aspects of electronic transport in nanoscale devices

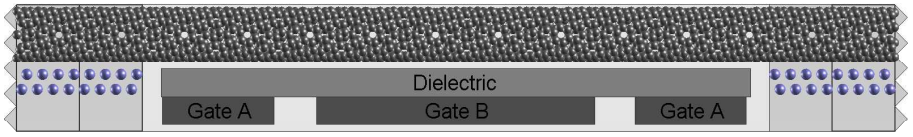
Sørensen, Hans Henrik Brandenborg

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Computational aspects of electronic transport in nanoscale devices

Hans Henrik Brandenborg Sørensen



Kongens Lyngby 2008
IMM-PHD-2008-195

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This thesis is concerned with the modeling of electronic properties of nano-scale devices. In particular the computational aspects of calculating the transmission and current-voltage characteristics of Landauer-Büttiker two-probe systems are in focus. To begin with, the main existing methods are described in detail and benchmarked. These are the Green's function method and the wave function matching method. The methods are subsequently combined in a hybrid scheme in order to benefit from a common formalism.

The most time demanding stages of common electronic transport calculations are identified. For systems of more than about a hundred atoms, two specific tasks stand out; the evaluation of self-energy matrices to describe the coupling between the electrodes and the device, and the solution of the central region Schrödinger equation either by matrix inverse of by solving a system of linear equations. In this work the objective is to develop new efficient algorithms for these tasks in order to model nano-scale systems of larger size in the future. The starting point of the new methods is the combined formalism of the Green's function and wave function matching methods.

The first new algorithm described is for the calculation of the block tridiagonal matrix inverse of a block tridiagonal matrix in $O(N)$ operations. This algorithm also leads to an optimal evaluation of the frequently used Caroli transmission formula. A modified wave function matching scheme is then developed which allows for a significant reduction in the cost of the self-energy matrix calculations when combined with an iterative eigensolver. Finally, such an iterative eigensolver is developed and implemented based of a shift-and-invert Krylov subspace approach. The method is applied to a selection of nano-scale systems and speed-ups of up to an order of magnitude are achieved.

Resumé

Denne afhandling omhandler modellering af de elektroniske egenskaber for komponenter af nano-størrelse. Specifikt de numeriske aspekter i at beregne transmission af elektroner og sammenhængen mellem strøm og spændingsforskel for Landauer-Büttiker to-elektrode-systemer er i fokus. Til at begynde med beskrives de vigtigste eksisterende metoder. Disse metoder kaldes for Green's-function-metoden og wave-function-matching-metoden. Det vises hvordan man sammenføjer formalismerne til en fordelagtig hybrid.

De mest tidskrævende skridt for en almindelig beregning af elektrontransport identificeres. For systemer med mere end ca 100 atomer, er det specielt to beregningstunge opgaver der træder frem; beregningen af selv-energi-matricer, der beskriver koblingen mellem elektroderne og komponenten i midten, og løsningen til Schrödingerligningen for komponentdelen, i form af en matrix-invertering eller en løsning af et linært system. I denne afhandling er målet, at udvikle nye og effektive algoritmer for disse to opgaver, med henblik på at kunne modellere større nano-systemer i fremtiden. Udgangspunktet for de nye algoritmer, er den kombinerede formalisme for Green's-function- og wave-function-matching-metoderne.

Den første ny algoritme der beskrives er til beregning af den blok-tridiagonale del af den inverse af en blok-tridiagonal matrix, hvilket gøres i $O(N)$ kompleksitet. Denne algoritme leder også direkte til en optimal udregning af transmissionen via Caroli's formel. Dernæst udvikles en modificeret wave-function-matching-metode, som giver anledning til betydeligt hurtigere beregninger af selv-energi-matricerne, hvis den kombineres med en iterativ egenværdiløser. Til sidst udvikles og implementeres en sådan egenværdiløser, baseret på en shift-and-invert Krylov underrumsmetode. Metoden anvendes på et udvalg af forskellige nano-systemer, hvorved der opnås besparelser i beregningstiderne på op til en størrelsesorden.

Preface

This thesis is submitted to the department of Informatics and Mathematical Modeling (IMM) at the Technical University of Denmark (DTU) in partial fulfillment of the requirements of the Ph.D. degree. It is based on work which was carried out in the Scientific Computing group, IMM, DTU, from April 2005 to April 2008 under the supervision of Professor Per Christian Hansen and Professor Kurt Stokbro. Research visits has been conducted at the Earth Simulator Center, RIST, Tokyo, Japan, and the Institute for Applied Mathematics, Delft University of Technology, Delft, The Netherlands.

I am truly grateful to many people who contributed to my work at DTU. First of all, I would like to acknowledge my primary supervisor, Per Christian Hansen, for his guidance and encouragement throughout my Ph.D. career. His open-minded attitude towards research created the free environment that has allowed me to investigate scientific subjects that truly excite me. His constant enthusiasm and outstanding skills gave me great encouragement and were greatly appreciated. I would also like to thank the other members of my supervisory group, Kurt Stokbro and Stig Skelboe, for the time and the insight they provided. I will miss the weekly lunch meetings and interesting discussions.

My gratitude also goes to current and former colleagues. I would especially like to thank Dan Erik Petersen-san, who has been my Ph.D.-mate and good friend for the last three years. I enjoyed many conversations with him about research and life. I would also like to thank Jesper Gross and Mark Hoffmann, who came before me and offered assistance when it was needed, and Bernd Drammann for his great teaching, helpfullness and friendship. My appreciation also goes to Martin van Gijzen who was my host and co-worker during my stay in Delft. I was lucky to have such a knowledgeable and hospitable host.

Finally, I would like to express my great appreciation to my family for their support, to my friend Christoffer Dam Bruun, for proof-reading and outstanding friendship, and last but not least, to my girlfriend, Tina Roden, for her love, understanding and patience.

Kgs. Lyngby, December 2008

Hans Henrik Brandenborg Sørensen

List of included papers

PAPER I

Block tridiagonal matrix inversion and fast transmission calculations
Dan Erik Petersen, Hans Henrik B. Sørensen, Per Christian Hansen, Stig Skelboe, Kurt Stokbro
Journal of Computational Physics 227, 3174 (2008)

PAPER II

Efficient wave function matching approach for quantum transport calculations
Hans Henrik B. Sørensen, Per Christian Hansen, Dan Erik Petersen, Stig Skelboe, Kurt Stokbro
Physical Review **B**, to be submitted

PAPER III

Krylov subspace method for evaluating the self-energy matrices in electron transport calculations
Hans Henrik B. Sørensen, Per Christian Hansen, Dan Erik Petersen, Stig Skelboe, Kurt Stokbro
Physical Review **B** 77, 155301 (2008)

Contents

Chapter 1

Introduction

Over the last few decades, the fundamental techniques for integrating elementary circuits on a chip have undergone dramatic advances. This has had a tremendous impact on the technology around us, our everyday lives, and on the computational sciences so that, for example, the calculations performed in the current thesis have become possible. Today, the feature size of integrated circuits are of the order of 100 nm. In the near future this order will most likely approach 10 nm and lead to yet more powerful computers. Unavoidably, the continuation of this trend, well known as Moore's law, will reach a domain where the feature size becomes comparable to the wave length of the electrons in the circuits. In that case, quantum mechanical effects can arise and the classical Ohm's law of electronic transport breaks down. On the other hand, such quantum effects are not necessarily harmful if we have the basic understanding of them, as they can be used to design radically new types of electronic components. The development and manufacturing of such components at the scale of nanometers is the key aspect and idea of nano-technology.

Besides technological prospects, nano-scale devices also offer a convenient platform to explore the fundamental physics of electrons in solids and molecules. In this ongoing journey of understanding the nature of materials, researchers around the world have developed many theories, among which the density functional theory (DFT) has emerged as one of the most powerful tools. This theory, which in its modern version dates back to the seminal work by Kohn and his collaborators in the mid-1960s, has been the most prominent contributer to the rising field of computational materials science. In this field, the basic properties of atomic configurations are investigated from first principles by using numerical simulations, which are often quite expensive. Consequently, such computational approaches depend equally on both state-of-the-art computer hardware and efficient numerical algorithms in order to be successful.

One of the most active areas of computational materials research in recent years, yet also one of the numerically most costly, is the combination of DFT with modern theories of quantum transport in the attempt to study electronic transport through nano-scale devices under external bias. Among the variety

of schemes developed, a special place is occupied by the application of the non-equilibrium Green's function (NEGF) method within the DFT approach. This particular approach is by far the most widely used method, and, with some appropriate approximations and optimized algorithms, NEGF-DFT has become a cost-effective first-principles method and has reproduced a variety of ground-state properties within a few percentage of error compared with experimental data. However, for the vast majority of interesting cases, the sizes of the nano-scale systems considered in calculations have been much smaller than realistic experimental configurations. Usually the number of atoms is no more than a few hundred. The reason for this is in general a lack of sufficient computational resources and/or efficiently implemented numerical algorithms.

In this thesis, we develop new computational methods for the modeling of the transport properties of nano-scale devices. The purpose of this is to treat systems with thousands of atoms and thus reduce the gab between simulations and realistic present day experiments. A typical example of such a large system, which is modeled in the current thesis, is shown on the cover page. It consists of a semi-conducting carbon nanotube (CNT) positioned on Lithium metallic surfaces at each end and next to an arrangement of three gates. This particular configuration is of interest as a field-effect transistor (FET). The entire system contains 1760 atoms and hence represents an immense computational task using standard first principles methods, even on a high performance computing platform. In order to obtain the electronic structure, current-voltage characteristics, and other important properties of such systems, we have implemented optimizations and novel algorithms for the computationally most expensive stages of the calculation. In the most favorable cases, the developed algorithms are up to an order of magnitude faster than existing methods.

The current thesis presents the work related to the development and computational aspects of our new algorithms, which has lead to the publication of three scientific papers. Specifically, my contributions to this work include:

- Formulation and implementation of a framework to combine the Green's function method and the wave function matching method in practical calculations (Chapter 3).
- Co-development of a new inversion algorithm to obtain the block tridiagonal Green's function matrix. Co-development of a faster transmission calculation in the Green's function method (Chapter 4/**PAPER 1**).
- Development and implementation of a modified wave function matching method which is computationally very efficient (Chapter 5/**PAPER 2**).
- Development and implementation of a Krylov subspace method for the calculation of self-energy matrices (Chapter 6/**PAPER 3**).
- The study of band-to-band tunneling in a carbon nanotube based FET device of 14 nm in length and more than 1500 atoms (Section 6.4).

1.1 Units, notation, and key linear algebra

Unless otherwise specified or explicitly not the case, we will use Hartree atomic units ($m_e = \hbar = e = 1$) throughout this text.

The general principles used for notation in this thesis is as follows. Scalars are written as ordinary Roman letters, e.g., $\mathrm{x}, \mathrm{y}, \mathrm{a}, \mathrm{A}$, etc. Integer variables are predominantly given in italic, i, j, k, etc.. Vectors are denoted by lower case bold face characters, $\mathbf{x}, \mathbf{y}, \mathbf{a}, \mathbf{b}$, etc., and matrices as bold face capital letters $\mathbf{X}, \mathbf{Y}, \mathbf{A}, \mathbf{B}$, etc. To index elements of vectors we use the notation $\mathrm{a}_i = [\mathbf{a}]_i$ and similarly $\mathrm{A}_{ii} = [\mathbf{A}]_{i,i}$ and $\mathrm{A}_{i,j} = [\mathbf{A}]_{i,j}$ for matrices. Finally, a particular block (i, j) of a matrix $\mathbf{A}$ that has block structure is denoted by $\mathbf{A}_{i,j}$ (see below).

We use several identities and relations linear algebra in this thesis. Two of these, which are of central importance and used repeatedly in most chapters, will be stated here for convenience:

- **Matrix inverse:** For any full-rank square matrix $\mathbf{A}$ with a 2×2 block structure,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{1.1}$$

  one can obtain the inverse matrix also of 2×2 block form, given by [1]

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{S}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{S}^{-1} \end{pmatrix}, \tag{1.2}$$

  where $\mathbf{S} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ is the so-called Schur complement block.

- **Block Gaussian elimination:** Generalizing the 2×2 block inverse operation to general block tri-diagonal form, eliminating the lower blocks of the augmented matrix $(\mathbf{A}|\mathbf{I})$ can be written [1]

$$\left(\begin{array}{cccc|cccc} \mathbf{A}_{11} & \mathbf{A}_{12} & & & \mathbf{I} & & & \\ \mathbf{A}_{21} & \ddots & \ddots & & & \mathbf{I} & & \\ & \ddots & \ddots & \mathbf{A}_{\mathrm{n-1,n}} & & & \ddots & \\ & & \mathbf{A}_{\mathrm{n-1,n}}^{\dagger} & \mathbf{A}_{\mathrm{n,n}} & & & & \mathbf{I} \end{array}\right) \tag{1.3}$$

$$\sim \left(\begin{array}{cccc|cccc} \mathbf{A}'_{11} & \mathbf{A}_{12} & & & \mathbf{I} & & & \\ & \mathbf{A}'_{22} & \ddots & & \mathbf{J}_{2,1} & \mathbf{I} & & \\ & & \ddots & \mathbf{A}_{\mathrm{n-1,n}} & \vdots & \ddots & \ddots & \\ & & & \mathbf{A}'_{\mathrm{n,n}} & \mathbf{J}_{\mathrm{n},1} & \cdots & \mathbf{J}_{\mathrm{n,n-1}} & \mathbf{I} \end{array}\right), \tag{1.4}$$

  where $\mathbf{A}'_{11} = \mathbf{A}_{11}$ and

$$\mathbf{A}'_{ii} = \mathbf{A}_{ii} - \mathbf{A}_{i,i-1}(\mathbf{A}'_{i-1,i-1})^{-1}\mathbf{A}_{i-1,i}, \quad i > 1 \tag{1.5}$$

$$\mathbf{J}_{ij} = -\mathbf{A}_{i,i-1}(\mathbf{A}'_{i-1,i-1})^{-1}\mathbf{J}_{i-1,j}, \quad i > 1, j > i, \tag{1.6}$$

which gives the inverse of the nth row as $[\mathbf{A}^{-1}]_{\mathrm{n},j} = \mathbf{J}_{\mathrm{n},j}$, $1 \leq j < \mathrm{n}$ and $[\mathbf{A}^{-1}]_{\mathrm{n,n}} = (\mathbf{A}'_{\mathrm{n,n}})^{-1}$.

In the following, we will refer to the second procedure listed here as a downwards block Gaussian elimination "sweep".

1.2 Outline

This thesis is divided into 7 chapters, 3 appendices and the bibliography. It is organized as follows. In Chap. 2, we introduce DFT which forms the starting point and theoretical foundation of our computational methods. In Chap. 3, we discuss the state-of-the-art modeling of electronic transport from a numerical point of view based on the Landauer picture of quantum transport in weakly interacting phase coherent systems. In Chap. 4, we describe possible computational optimizations within the Green's function method. In Chap. 5, we develop the formalism for a more efficient implementation of the wave function matching method. In Chap. 6, we describe a Krylov subspace algorithm for the very fast evaluation of self-energy matrices within the formalism of the previous chapter. In Chap. 7, we give a short conclusion and outlook.

Chapter 2

Preliminary concepts: elements from solid state physics

In order to make the reader minimally familiar with the important concepts and approximations that underly the numerical calculation of electronic transport, we will begin this thesis from the point of view of solid state physicists. The goal is to convey in few words how the stationary quantum mechanical properties of nano-scale structures can be found from the self-consitent solution of an eigenvalue problem and a Poisson problem. In addition, we also wish to demonstrate that the solution can be obtained in a numerical calculation on a standard computer, in a matter of minutes, if the structures are relatively small or bulk-like. We refer the reader to the excellent text books by Ashcroft and Mermin [2], Martin [3], and Datta [4] for more detailed presentations.

2.1 Electronic structure calculations

Nano-scale devices usually consist of thousands of atoms which, in turn, contains equally many nuclei and possibly a much larger number of electrons. Nuclei and electrons interact with each other and mutually among themselves. These interactions establishes the structure and electronic configuration of the system and must be properly described by the many-particle Schrödinger's equation

$$\hat{H}\Psi(\mathbf{r}_1,\ldots,\mathbf{r}_N;\{\mathbf{R}_m\}) = E\Psi(\mathbf{r}_1,\ldots,\mathbf{r}_N;\{\mathbf{R}_m\}), \tag{2.1}$$

where $\mathbf{r}_i$ are the positions of the N electrons, $\{\mathbf{R}_m\}$ is the set of nuclei coordinates which is assumed fixed in space (Born-Oppenheimer approximation), and the many-particle Hamiltonian operator is defined as

$$\hat{H} = -\frac{1}{2}\sum_i \nabla_i^2 + \sum_{i<j}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} + \sum_i V_{\text{ext}}(\mathbf{r}_i;\{\mathbf{R}_m\}), \tag{2.2}$$

composed of terms for the kinetic energy, the electron-electron interaction, and an external potential, which in this description accommodates the influence of the electron-nucleus and nucleus-nucleus interactions. The nuclei equilibrium positions can be found by minimizing the total energy E with respect to $\{\mathbf{R}_m\}$.

2.1.1 Density-functional theory

An unfortunate aspect of the above formulation is that the many-particle wave function Ψ provides the complete knowledge of the stationary properties of the system, but is numerically infeasible to compute for more than a few particles [5]. It is therefore more practical to consider the one-body electron density,

$$n(\mathbf{r}) = N \int \mathrm{d}\mathbf{r}_2 \cdots \int \mathrm{d}\mathbf{r}_N |\Psi(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N; \{\mathbf{R}_m\})|^2, \tag{2.3}$$

as the central quantity of interest. The number of degrees of freedom is then reduced from $3N$ to 3 if Ψ is not explicitly evaluated. Moreover, the Hohenberg-Kohn theorems [6] prove that *all* ground state properties of the many-particle system can be calculated from the unique ground state density $n(\mathbf{r})$ that corresponds to the non-degenerate ground state energy $E_0 = \min E[n(\mathbf{r})]$.

2.1.2 Kohn-Sham equations

Kohn and Sham [7, 8] came up with a clever approach to minimize the energy functional $E[n(\mathbf{r})]$ and obtain the desired ground state density $n(\mathbf{r})$ by solving a set of one-particle eigenvalue equations, given by

$$\Big[-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}) \Big] \psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}), \tag{2.4}$$

$$v_{\mathrm{eff}}(\mathbf{r}) = V_{\mathrm{H}}(\mathbf{r}) + V_{\mathrm{xc}}(\mathbf{r}) + V_{\mathrm{ext}}(\mathbf{r}), \tag{2.5}$$

$$n(\mathbf{r}) = \sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2, \tag{2.6}$$

$$V_{\mathrm{H}}(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, \mathrm{d}\mathbf{r}', \quad V_{\mathrm{xc}}(\mathbf{r}) = \frac{\delta E_{xc}[n(\mathbf{r})]}{\delta n(\mathbf{r})}, \tag{2.7}$$

in a self-consistent manner. The trick is to split the complicated electron-electron interactions into the averaged effect of the other electrons (the Hartree potential term $V_{\mathrm{H}}(\mathbf{r})$ of v_{eff}) and the rest (the exchange-correlation (XC) term $V_{\mathrm{xc}}(\mathbf{r})$ of v_{eff}). Note also that the eigenfunctions $\psi_i(\mathbf{r})$ and energies ϵ_i in these equations are merely auxiliary and do not have any strict physical meaning [8].

2.1.3 Local density approximation

The Kohn-Sham self-consistent procedure is remarkably efficient from a computational point of view [3]. To be exact, however, it requires the correct exchange-correlation functional $E_{xc}[n(\mathbf{r})]$ and this is not available in practice. Additional approximations has to be done, for which there exists a variety of schemes [9] with details beyond the scope of this short introduction. Here we mention only the simplest, the local density approximation (LDA), which is used in all calculations of the current work. With this approximation, the exchange-correlation energy per electron for a uniform electron gas, $\epsilon_{xc}^{n_0} \equiv \frac{E_{xc}(n_0)}{n_0}$, is extrapolated to an inhomogeneous distribution, i.e,

$$E_{xc}^{\mathrm{LDA}}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{xc}^{n_0}\big(n(\mathbf{r})\big)\,\mathrm{d}\mathbf{r}. \tag{2.8}$$

Efficient parameterizations have been developed for the E_{xc}^{LDA} functional [10, 11].

2.1.4 Pseudopotential approximation

The main difficulty in electronic structure calculations is to treat the effects of the electron-electron interactions. However, since the inner-shell electrons, localized around each atom, often interact only weakly with the inner-shell electrons of other atoms, it is convenient to assume these electrons to be frozen. As such, they act much like the nuclei in being treated as an external potential so that we have the Kohn-Sham equations for the valence electrons alone.

Although this leads to a significant reduction in the number of electrons to be handled, another problem remains: The valence states $\psi_i(\mathbf{r})$ are rapidly varying in the core region, which is hard to represent numerically. In order to improve on this, the external potential for the nuclei plus core electrons (terms in V_{ext}) may be replaced by a smooth, non-singular potential known as a pseudopotential.

The most frequently used pseudopotential is the norm-conserved pseudopotential developed by Hamann, Schlüter and Chiang [12, 13]. The pseudo-wave functions $\psi_i^{\mathrm{ps}}(\mathbf{r})$ that are obtained with these pseudopotentials have no nodes and coincides with the correct $\psi_i(\mathbf{r})$ outside the inner-shell radius r_c. In addition, the pseudopotential eigenvalues ϵ_i^{ps} are made to agree with the all-electron eigenvalues ϵ_i, and the charge inside r_c (i.e, the norm $\int_0^{r_c} r^2|\varphi_l(r)|^2\,\mathrm{d}r$, where $\varphi_l(r)$ is the radial component of $\psi_i(\mathbf{r})$) is conserved. Efficient parameterization methods have been derived by Troullier and Martins [14].

2.1.5 Basis sets

The final requisite in order to solve the Kohn-Sham equations in practice is to represent the electronic wave functions $\psi_i(\mathbf{r})$ (here the "ps" superscript is implied) in an efficient manner. In general, one chooses between a real space grid, a plane wave basis, or a real space basis. Several factors decide the best

choice of basis, including ease of implementation, accuracy, and the type of system [3].

In grid-based methods, the wave function is represented on a mesh of points in real space, which makes them inherently simple to program, as many operations are easier in real space. For example, if $\psi_i(\mathbf{r})$ is explicitly given on a grid, then the calculation of the density in Eq. (2.6) is simply summing the squares at each point. It is also possible to employ the wide selection of well-established numerical methods for second-order differential equations: finite element, finite difference, multigrid, wavelets, etc. The accuracy is then easily set by means of the grid spacing. As grid-based methods are not used in this work, we will not consider this further. A recent review has been written by Beck [15].

The plane-wave basis, e.g., $\psi_i(\mathbf{r}) = \sum_{\mathbf{k}} c_{i,\mathbf{k}} e^{-\imath \mathbf{k}\cdot\mathbf{r}}$, where $\mathbf{k}$ are the wave vectors, is localized in reciprocal $\mathbf{k}$-space, and is ideal for studying crystals or other systems with periodic boundary conditions because the sum in the basis expansion can be limited to a set of reciprocal lattice vectors [2]. It is then fairly simple to implement and has an accuracy which can be controlled in a systematic fashion by specifying a cutoff value $|\mathbf{k}| < k_{\text{cutoff}}$. However, typically a relatively large number of basis functions is needed to achieve good accuracy [16]. Although the plane-wave basis is not put to practical use in our calculations, we will revisit the formalism in the discussion on $\mathbf{k}$-point sampling in Sec. 2.1.7.

In our computations, we will use a basis of functions $\{\phi_j\}$ which are localized in real space, i.e., on each atom as $\psi_i(\mathbf{r}) = \sum_j c_{i,j}\phi_j(\mathbf{r} - \mathbf{R}_m)$ for atom m. This choice gives us the freedom to select the basis functions so as to look much like our expected wave function, meaning, in principle, that we can represent the wave function accurately with just a few terms. Many choices are available, most based on Gaussians or discretized functions. In our case, we will employ linear combinations of atomic orbitals (LCAO), which can be written as $\phi_{lmn}(\mathbf{r}) = \phi_{ln}(r)Y_{lm}(\hat{\mathbf{r}})$, where Y_{lm} are spherical harmonics and $\phi_{ln}(r)$ is the radial dependence, commonly tabulated and stored on disk [17]. In general, there will be several orbitals (labeled by n) with the same angular momentum (labeled by l, m), but different radial dependence. This is conventionally called a multiple-ζ basis [3]. We will only use the single–ζ (SZ) and, on a few occasions, the double–ζ (DZ) basis in the numerical examples of this thesis.

2.1.6 Localization

It is important to mention, that the real space basis functions $\phi_{lmn}(\mathbf{r})$ are given only local support, i.e., they are zero beyond a certain radius. This allows for sparse representations of the Hamiltonian and overlap matrices (see Sec. 3.1.4) since the overlap between orbitals is limited to a few neighboring atoms. Such an approximation is well founded in the general principle of nearsightedness, as coined by Kohn [18], and makes it possible to compute the Hamiltonian and subsequently solve the Kohn-Sham equations in $O(N_{\text{basis}})$ operations, so-called linear scaling [19]. The drawback is, that these methods are relatively difficult

to implement and there is no systematic procedure to control the accuracy [3].

We note that the DFT scheme in the ATK program used as part of this thesis, is *not* a linear scaling implementation [20]. However, we still benefit highly from localization and the subsequent structure of the Hamiltonian in our calculations of electronic transport, see Chap. 3.

2.1.7 k-point sampling and bands

A final aspect of the KS equations is of fundamental importance: If the potential $v_{\text{eff}}(\mathbf{r})$ is periodic, then the solutions ψ_i to the KS eigenproblem in Eq. (2.4) can be expressed as so-called Bloch wave functions. This is for example the case for the bulk electrodes considered in this thesis, having translational symmetry $v_{\text{eff}}(\mathbf{r}) = v_{\text{eff}}(\mathbf{r}+\mathbf{T})$ for some spatial vector $\mathbf{T}$ in the direction of transport.

A Bloch wave, or Bloch mode, consists of the product of a plane wave envelope function $e^{\hat{\imath}\mathbf{k}\cdot\mathbf{r}}$ and a periodic function, written as

$$\psi_{i,\mathbf{k}}(\mathbf{r}) = e^{\hat{\imath}\mathbf{k}\cdot\mathbf{r}} u_{i,\mathbf{k}}(\mathbf{r}), \tag{2.9}$$

where $\mathbf{k}$ is the wave vector and $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}+\mathbf{T})$ has the same periodicity as the potential. The result that the eigenfunctions can be written in this form is called Bloch's theorem [2]. Inserting Eq. (2.9) into Eq. (2.4) allows us to write

$$\hat{\mathrm{H}}^{KS}(\mathbf{k}) u_{i,\mathbf{k}}(\mathbf{r}) = \epsilon_{i,\mathbf{k}} u_{i,\mathbf{k}}(\mathbf{r}), \tag{2.10}$$

where the Kohn-Sham Hamiltonian operator

$$\hat{\mathrm{H}}^{KS}(\mathbf{k}) = e^{-\hat{\imath}\mathbf{k}\cdot\mathbf{r}}\Big[-\frac{1}{2}\nabla^2 + v_{\text{eff}}(\mathbf{r})\Big]e^{\hat{\imath}\mathbf{k}\cdot\mathbf{r}} = -\frac{1}{2}(\nabla + \hat{\imath}\mathbf{k})^2 + v_{\text{eff}}(\mathbf{r}) \tag{2.11}$$

becomes explicitly $\mathbf{k}$ dependent. Furthermore, since $u_{\mathbf{k}}(\mathbf{r})$ is periodic, one can arive at the Fourier series

$$\psi_{i,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}} e^{\hat{\imath}(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \tag{2.12}$$

where the sum is over vectors $\mathbf{G}$ belonging to the reciprocal bulk lattice, the dual of the real-space lattice [3]. Although $\mathbf{k}$ is arbitrary, we see that adding a reciprocal lattice vector $\mathbf{G}$ to $\mathbf{k}$ simply shuffles the sum in Eq. (2.12). As a result, $\psi_{i,\mathbf{k}}(\mathbf{r})$ can always be characterized by a wave vector $\mathbf{k}$ inside a "unit cell" of the reciprocal lattice, the so-called Brillouin zone (BZ). The corresponding energy eigenvalues $\epsilon_{i,\mathbf{k}}$ may also be found separately for each $\mathbf{k}$. Because the discrete energies associated with the indices i vary continuously with $\mathbf{k}$ we speak of an energy band (or energy gab where there are no solutions for any $\mathbf{k}$).

From the expressions above we see that the Hamiltonian and other $\mathbf{k}$ dependent quantities may be obtained as an average value "per unit cell" from an integral of the form [3]

$$\bar{f} = \frac{1}{\Omega_{\text{BZ}}} \int_{\text{BZ}} \mathrm{d}\mathbf{k} f(\mathbf{k}), \tag{2.13}$$

where f is a general function of $\mathbf{k}$. Since adjacent $\mathbf{k}$-points results in almost identical $f(\mathbf{k})$ values, the integral can be computed by summing values for the integrand at a limited number of $\mathbf{k}$-points in the BZ. Using symmetry it is easy to show that a number of $\mathbf{k}$-points are equivalent (high-symmetry points) and this reduces the number of required evaluations of $f(\mathbf{k})$ even further. We end up with a finite sum over a grid of $\mathbf{k}$-points,

$$\bar{f} = \sum_{\mathbf{k}} w_{\mathbf{k}} f(\mathbf{k}), \tag{2.14}$$

where $w_{\mathbf{k}}$ are appropriate weights. The optimal choice of the $\mathbf{k}$-points or sampling is often found as a balance between accuracy and efficiency. In general, for metallic systems the $\mathbf{k}$-point grids should be quite dense due to a complex shape of the Fermi surface in metals [3]. In contrast, insulators and semi-conductors require relatively few $\mathbf{k}$-points to sample.

In the calculations of this work, we will employ the $\mathbf{k}$-point sampling scheme developed by Monkhorst and Pack [21]. This scheme applies a regular grid as $(N_{\hat{\mathbf{x}}}, N_{\hat{\mathbf{y}}}, N_{\hat{\mathbf{z}}})$ shifted by one-half of the grid spacing. At times we also adopt the so-called Γ-point approximation (see, e.g., the transmission calculations in Sec. 3.5) for which only the $\mathbf{k} = \mathbf{0}$ point of the BZ is considered. For details on the validity of this approximation see for example Ref. [22].

2.2 Numerical implementation: the atk program

We will end this introductory chapter by discussing the necessary computational steps needed for the numerical solution of the KS equations (2.4)–(2.6). At the same time this serves as a description of the implementation in the ATK program.

2.2.1 Self-consistent procedure

It is apparent from the KS expressions, that the eigenvalue equations in Eq. (2.4) depend on their own solutions $\{\psi_i\}$ through the effective potential that depends on the density $n(\mathbf{r})$. This then corresponds to a non-linear eigenvalue problem which must be solved self-consistently. An appropriate flow-diagram to do this is presented in Fig. 2.1 in the manner it is implemented in the ATK program.

We now briefly outline each step of this self-consistent procedure (a detailed description can be found on the official website [23]). The key computational objects stored in an iteration are the density $n(\mathbf{r})$, the Hamiltonian matrix $\mathbf{H}$, and the single-particle wave functions ψ_i. From the outset, the representations of these objects are given in the LCAO basis $\{\phi_j\}$ mentioned above (we use compact index notation $j \equiv \{Ilmn\}$ to designate the basis orbitals). Subsequently, they will be given by relatively small $\sim O(N_{\mathrm{el}})$ matrices and vectors, labeled as $\mathbf{n}$, $\mathbf{H}$, and $\mathbf{c}_i$, respectively, in the following.

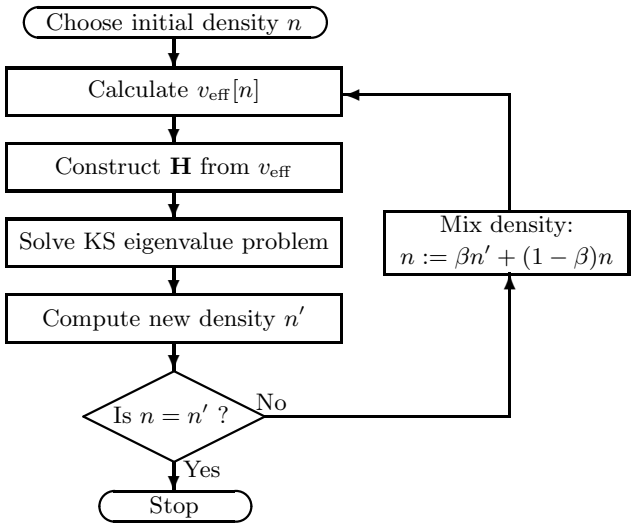


Figure 2.1: Schematic flow-diagram for the self-consistent KS procedure.

Initial density

It is common to begin the self-consistent procedure by specifying the initial density. If no *a priori* knowledge is available about the final density, a natural choice is simply to sum the densities of neutral atoms: $n(\mathbf{r}) = \sum_m n^{\text{atom}}(\mathbf{r} - \mathbf{R}_m)$. Alternatively, a better guess at the final density may be available based on a previous solution of the KS equations. In particular, if the atomic configuration is unchanged (e.g, if only the device voltage or gate voltage settings are different, as is the case in Sec. 5.7) with respect to an earlier result, then there is much to gain from storing and reusing the electronic densities.

The ATK program converges if possible no matter what the quality of the initial density, however, the steps required are much less with a good guess.

Calculate $v_{\text{eff}}(\mathbf{r})$ from the density

The first term in Eq. (2.5), the Hartree potential $V_{\text{H}}(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, \mathrm{d}\mathbf{r}'$, is much too expensive to compute by direct numerical integration. Therefore, we evaluate it by solving the Poisson equation $\nabla^2 V_{\text{H}}(\mathbf{r}) = -4\pi n(\mathbf{r})$ on a uniform spatial grid with suitable boundary conditions on the surface S. From the pool of techniques [1] to solve the Poisson equation, we employ the multigrid (MG) method [24] and the fast Fourier transform (FFT) [25], depending on the boundary conditions. For bulk systems and supercell systems large enough not to interact with its images, we may use periodic boundary conditions and the FFT.

Otherwise, we use MG when Dirichlet ($V_{\mathrm{H}}|_S = 0$) or Neumann ($\nabla V_{\mathrm{H}}|_S = 0$) boundary conditions are specified. It is possible to select appropriate boundary conditions separately for each spatial dimension of the Poisson problem. The details of having the proper boundary conditions for open transport problems is elaborately described in Ref. [26].

The calculation of the XC term $V_{\mathrm{xc}}(\mathbf{r})$ can be implemented according to the parameterization listed in App. C of Perdew and Zunger [11], i.e., its value at a specific point $\mathbf{r}$ is computed from a short polynomial expression in $r_s = (\frac{3}{4\pi n(\mathbf{r})})^{1/3}$ having different coefficients for $r_s \geq 1$ (low density limit) and $r_s < 1$ (high density limit). Again, it suffices to do this for a uniform spatial grid.

Last, we need to evaluate the norm-conserved pseudopotential $V_{\mathrm{ps}}(\mathbf{r})$. We use the non-local Kleinman-Bylander (KB) form [27] in the way devised by the SIESTA method [17]. With this approach the pseudopotential is separated into local and non-local parts $V_{\mathrm{ps}}(\mathbf{r}) = \sum_I V_I^{\mathrm{ps,local}}(\mathbf{r} - \mathbf{R}_I) + \sum_I \hat{V}_I^{\mathrm{ps,KB}}$, where $\hat{V}_I^{\mathrm{ps,KB}}$ is the KB projection operator. Tables for efficient calculations have been pretabulated and are widely available on the internet (ATK reads the unified pseudopotential file (UPF) format) [28]. Since $V_{\mathrm{ps}}(\mathbf{r})$ is independent on n and unchanged throughout the self-consistent loop, it need only be evaluated once.

Constructing the Hamiltonian matrix

Expanding the KS wave functions ψ_i in the non-orthogonal LCAO basis $\{\phi_j\}$, turns Eq. (2.4) into a generalized eigenvalue problem $\mathbf{Hc} = E\mathbf{Sc}$, where the Hamiltonian and overlap matrix elements are given by the integrals

$$H_{ij} = \int \mathrm{d}\mathbf{r}\phi_i^*(\mathbf{r} - \mathbf{R}_I)\Big[-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r})\Big]\phi_j(\mathbf{r} - \mathbf{R}_J), \tag{2.15}$$

$$S_{ij} = \int \mathrm{d}\mathbf{r}\phi_i^*(\mathbf{r} - \mathbf{R}_I)\phi_j(\mathbf{r} - \mathbf{R}_J), \tag{2.16}$$

and $\mathbf{c}$ are the eigenvectors holding the expansion coefficients. Here orbitals i and j are assumed to be localized on atoms I and J, respectively (we use the notation ϕ_i^* for generality, despite the fact that our basis functions are real in practice). Moreover, the effective potential v_{eff} contains both local and non-local pseudopotential terms given in the KB form above. One should employ specialized techniques to evaluate the matrix element integrals in an efficient manner, as found in the literature [3, 17].

Solve KS eigenvalue problem

For small systems, $\mathbf{H}$ and $\mathbf{S}$ will be small, dense matrices of size $N \times N$, where N is the total number of orbitals in the LCAO basis. In such cases it is appropriate to employ traditional dense eigensolvers to solve $\mathbf{Hc} = E\mathbf{Sc}$, e.g., the LAPACK

routines DSYGV/ZHEGV [29].[1] This is in contrast to the plane-wave approaches which gives large-scale sparse eigenproblems that are more favorably solved by iterative methods [16]. For larger systems, one would expect a certain level of sparsity and block structure also in the minimal LCAO basis (see Sec. 3.1.4), but this is not utilized in the ATK eigensolver (although this can be done, see [19] and references therein). The reason is that all main systems considered in this thesis are two-probe systems setup for transport calculations. As shown in the next chapter, solving the KS eigenvalue problem in these cases is done by splitting it into three parts, two small periodic electrode problems, which are efficiently solved by the dense eigensolver, and one large central problem, solved as a Green's function problem, i.e., $\mathbf{G} = (\mathbf{H} - \epsilon\mathbf{S})^{-1}$. We show in Chap. 3 how to utilize the block sparsity of $\mathbf{H}$ and $\mathbf{S}$ for the matrix inverse in this formalism.

Compute the new density

From the KS eigenfunctions $\mathbf{c}_i$ we are able to evaluate the corresponding ground state density by filling the N_{el} lowest states: $n(\mathbf{r}) = 2\sum_{i=1}^{N} n_i|\psi_i(\mathbf{r})|^2$, where n_i are occupation numbers (0 for empty states, and 1 for filled states) and the factor of two accounts for spin degeneracy. In practice, the calculation is performed by first computing the density matrix $\mathbf{D}$, defined as

$$\mathrm{D}_{ij} = \mathbf{c}_j \mathrm{diag}\{n_1, \ldots, n_N\}\mathbf{c}_i^{\dagger}, \tag{2.17}$$

and then the new electronic density on a grid from the expression

$$n(\mathbf{r}) = 2\sum_{i,j=1}^{N} \phi_i^*(\mathbf{r} - \mathbf{R}_I)\mathrm{D}_{ij}\phi_j(\mathbf{r} - \mathbf{R}_J). \tag{2.18}$$

Since only a small number of basis orbitals are nonzero at a given grid point, the calculation of the density in Eq. (2.18) can be performed in $O(N)$ operations, once $\mathbf{D}$ is known. The weighted outer product in Eq. (2.17), however, has complexity $O(N^3)$. One therefore has to use special techniques which takes localization and the nearsightedness principle into account to make this step linearly scaling [17, 19]. Again, ATK does not implement order-N algorithms for small molecular and bulk systems, and we will not discuss them further here.

Achieving self-consistency

The previous step completes the self-consistent cycle. In practice, the procedure then represents a fixed point iteration $\mathcal{F}$ to find the electronic density n such that $n = \mathcal{F}[n]$. Unfortunately, if we directly apply the new density n' from one iteration as input in the next, the procedure will be unstable. The simplest

[1] DSYGV is not applicable in a general setting since two-probe systems and/or $\mathbf{k}$-point sampling produces complex Hamiltonian matrices, see Sec. 2.1.7.

cure for this is to mix the previous density n with only a fraction β of the new density n' to produce the input for the next iteration: $n := \beta n' + (1-\beta)n$.

The simple linear mixing is reasonable stable as long as only a very small β is used. However, this makes the convergence quite slow [3]. We will therefore apply the more sophisticated mixing scheme developed by Pulay [30], which generally accelerates the convergence quite significantly, and can reach convergence in cases where linear mixing cannot. The input density for the $k+1$th iteration is constructed using the input and output densities of a number K of former cycles, implemented in the following way:

$$n_{\text{in}}^{(k+1)} = \beta n_{\text{out}}^{(k)} + (1-\beta) n_{\text{in}}^{(k)}, \tag{2.19}$$

$$n_{\text{in}}^{(k)} = \sum_{i=1}^{K} \alpha_i n_{\text{in}}^{(k-K+i)}, \quad n_{\text{out}}^{(k)} = \sum_{i=1}^{K} \alpha_i n_{\text{out}}^{(k-K+i)}, \tag{2.20}$$

where the values of α_i are obtained by consecutively minimizing the distance between $n_{\text{in}}^{(k)}$ and $n_{\text{out}}^{(k)}$. Most self-consistent procedures usually converges faster and more smoothly if more previous cycles are used in the mixing. Using K of the order 10 to 20 is generally recommended [23], in particular for systems that converge poorly. In addition, finding the optimal value of the fraction β for a particular system can be tricky, and is largely based on experience. For electrically insulated systems, larger values of the mixing parameter can be used (up to 0.5). For metallic systems, the value should usually be around 0.1, whereas two-probe systems in particular might require a lower value (0.01) [23].

2.2.2 Benchmarks

To round off the discussion of the numerical implementation, let us look at some simple benchmark runs for the ATK program. These benchmarks can give an impression of actual time required for a DFT simulation and the computational expense of the different steps of the self-consistent procedure.

As mentioned, we will apply the self-consistent cycle illustrated in Fig. 2.1 in this work only for the relatively small electrode parts of two-probe systems. A modified cycle, described in the next chapter, is used for the larger central region parts. We here consider the relatively small example systems shown in Fig. 2.2, corresponding to three different hydrogenated diamond molecules, three different semi-conducting carbon nanotubes, both of growing sizes, and three different metals. In all cases, we represent the computational objects in a minimal single–ζ (SZ) basis, except for the gold electrodes, where single–ζ–polarized (DZP) is used. We apply periodic boundary conditions by taking advantage of the supercell method. For the molecules, this means that the simulation box has vacuum regions big enough in all dimensions to neglect interaction between repeated images. For the CNTs and metals, this is equally the case for the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ dimensions, while for the $\hat{\mathbf{z}}$ direction the system is

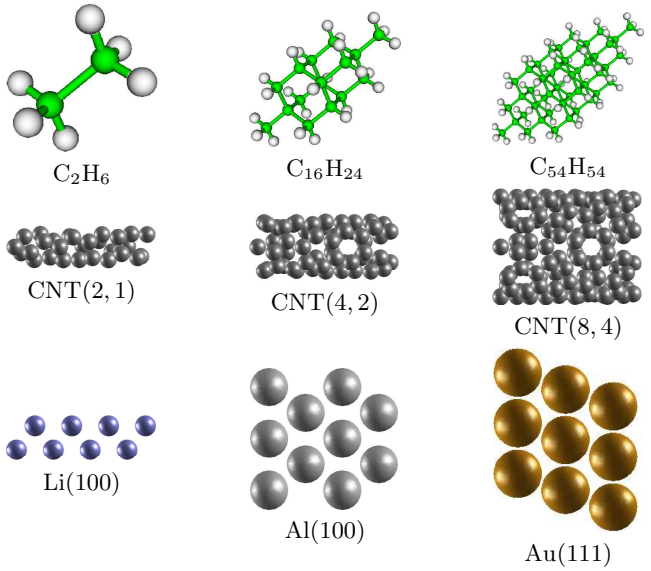


Figure 2.2: Example systems: Hydrogenated diamond molecules C_2H_6, $C_{16}H_{24}$, and $C_{54}H_{54}$; semi-conducting CNT$(2,1)$, CNT$(4,2)$, and CNT$(8,4)$ electrodes; bcc-lithium, fcc-aluminium, and fcc-gold metallic electrodes.

Table 2.1: CPU-times in seconds for calculating the self-consistent electronic density with the ATK program for the example systems given in Fig. 2.2. The system type, the **k**-point sampling of the Monkhorst type, the number of atoms, and the number of iterations are indicated in columns 2-5.

System	Type	**k**-points	Atoms	Iterations	CPU
C_2H_6	Molecule	-	8	11	64.2
$C_{16}H_{24}$	Molecule	-	40	12	701.4
$C_{54}H_{54}$	Molecule	-	108	11	2675.6
CNT(2,1)	Semi-conductor	(1,1,100)	28	11	183.5
CNT(4,2)	Semi-conductor	(1,1,100)	56	8	448.0
CNT(8,4)	Semi-conductor	(1,1,100)	112	8	2103.7
Li(100)	Metal	(1,1,100)	8	3	39.3
Al(100)	Metal	(1,1,100)	18	10	198.0
Au(111)	Metal	(1,1,100)	27	16	1357.6

periodic from the outset (for an accurate description we use 100 **k**-points along $\hat{\mathbf{z}}$ to sample the Brillouin zone). The Pulay mixing scheme is applied with $\beta = 0.1$ and 6 previous history steps taken into account.

Our benchmarking results for obtaining a converged density and Hamiltonian matrix are displayed in Table 2.1. The measured total CPU-times are listed in the last column. Also the number of iterations of the self-consistent procedure, the number of atoms treated, and the **k**-point samplings are indicated for each system. All systems converge to a tolerance of $||n - n'||_2 < 10^{-4}\text{Rydberg/Bohr}^3$ in 3–16 iterations. Even the most time consuming calculations take less than 45 minutes on a single CPU (we do not parallelize over **k**-points in these benchmarks, although this is possible) and will be insignificant in comparison with the total expense of the two-probe systems considered later in this thesis.

Still, it is quite informative to consider the current benchmark calculations in more detail. Therefore the CPU-times per iteration for the individual steps of the self-consistent procedure for the first six of the benchmark systems are presented in a semi-logarithmic plot in Fig. 2.3. We will comment here only on the main conclusions to be drawn:

- First, we can see that the initial density and Pulay mixing steps takes negligible time in all calculations.
- Second, it is apparent that the calculation of v_{eff} and, in particular, the $O(N\log N)$ scaling of the FFT Poisson solver, is the most costly step for the molecules, but with the construction of $\mathbf{H}$ a close second.
- Third, since the eigenvalue solution and the computation of the new density have the highest computational complexity $O(N^3)$, they tend to be-

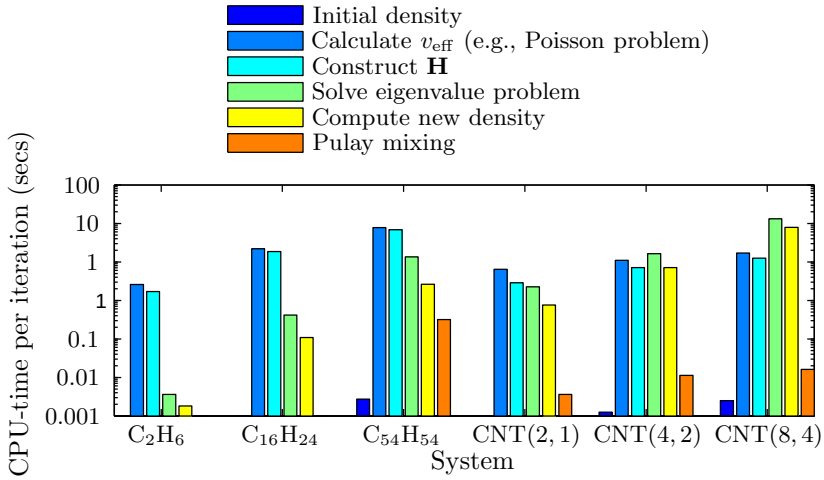


Figure 2.3: Benchmark results for the self-consistent cycle implemented in the ATK program. The measured CPU-times per iteration for the different steps of the self-consistent cycle are shown on a logarithmic axis.

come more and more dominating as the number of atoms increases. For larger systems, these steps will quickly become the overall bottle-neck (see Ref. [20]). This is already the case for the CNT systems because they are sampled at many **k**-points in each cycle.

All in all, we can conclude that the ATK program scales as $O(N\log N)$ for small molecules ($N <\sim 100$) and as $O(N^3)$ for large molecules and **k**-point sampled periodic bulk/electrode systems.

Chapter 3

Modeling of quantum transport through nano-scale devices

The purpose of this chapter is to present the computational aspects of modeling electronic transport in nano-scale devices. From a physical starting point, an accurate quantum mechanical model of these devices should have the ability to capture the following fundamental effects [4]:

- Electron wave interference.
- Quantum mechanical tunneling.
- Discrete energy levels due to confinement in 2D and 3D device geometries.
- Scattering processes (electron-phonon, electron-electron).

The first three effects can be accommodated most simply in the Landauer-Büttiker transport theory in conjunction with the solution of the electronic Schrödinger's equation, which was discussed in Chap. 2. Therefore, we begin the following presentation with a discussion of the phase-coherent[1] Landauer-Büttiker formalism. The fourth effect requires, in general, the NEGF formalism, or equivalent theories, to account for energy, momentum and phase relaxation due to inelastic scattering processes. However, if we are interested in electronic transport at low temperatures and apply only a small voltage across the device, the effects of inelastic scattering at the nano-scale can be neglected.

For simplicity, and since it also reflects the conditions of many experiments well, we will consider only this low-temperature-low-bias regime. In our case, the use of either the Landauer-Büttiker approach or the NEGF method then

[1] The terminology "phase-coherent" refers to a deterministic evolution of both the amplitude and phase of $\psi_{i,\mathbf{k}}(\mathbf{r})$ as given in Eq. (2.9). As explained below, the electron wave function evolves phase-coherently only in the case of elastic scattering.

becomes a matter of taste, since they are consequently equivalent. We will present both methods in this chapter, first the Green's function method relying of Caroli's formula for the transmission coefficient, and subsequently the wave function matching method, which explicitly determines the transmission and reflection probabilities of the Landauer-Büttiker theory. We will also combine the two methods in an attempt to make a fast hybrid approach. The chapter ends with benchmark timings of all the methods and a short comparison discussion.

3.1 Landauer-Büttiker formalism

A transparent quantum mechanical formulation of the electronic transport in small conductors was first proposed in 1957 by Landauer [31], who suggested a simple formula which established the relation between the transmission probability of the electron and the electronic conductance in one-dimensional structures. Landauer's idea was later generalized by Büttiker [32] to the case of multi-channel-multi-probe devices. Since then, most theoretical work on electronic transport at the nano-scale has been based on this formalism (see Ref. [4] and references therein) and it also represents the theoretical foundation for the computational methods used in the current thesis.

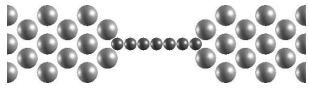
3.1.1 Phase-coherent transport of electrons

If one applies an electric field in the form of a potential difference V to a metal or a semi-conductor, the electrons in the vicinity of the Fermi level will be accelerated. While moving, the electrons scatter from possible impurities or lattice vibrations (phonons), and, if the scattering is inelastic, they may loose some of their acquired momentum. As a result, when the material is long enough, the process of repeated acceleration and inelastic scattering carries the electrons in the direction of the electric field with an equilibrium drift speed, producing a drift current I. The equation that describes this situation is the well-known Ohm's law: $V = RI$, where R is the resistance caused by the scatterers, which is proportional to the length of the device. The law governs for example the case of the standard resistor shown in the left part of Fig. 3.1.

In this thesis we consider very small devices, where it can be assumed that there are no impurity scatterers. More precisely, the mean free path that an electron moves on average before it has lost its original momentum ($10 - 50$ nm for metallic bulks at room temperature [33]) is longer than, or at the same scale as, the length of the device, see right part of Fig. 3.1. The electrons can then penetrate a bulk system without being scattered, and the electric resistance becomes independent of the length of a material. This type of conduction is called ballistic transport of electrons.

Unfortunately, since we will not deal with ideal bulk conductors only, but rather a device of a given atomic configuration, the transport will not be per-

~ 1 cm



~ 1 nm

Figure 3.1: Macroscopic and nano-scale electronic devices: a standard 2.6 MΩ resistor and the Al–C_7–Al nano-scale analogue.

fectly ballistic. In our nano-scale systems, quantum mechanical length scales have to be taken into account, i.e., the de Broglie wave length at the Fermi level ("electron wave length") and the phase-relaxation length. The latter is the average length an electron wave can propagate before inelastic collisions destroy its original phase. It is typically longer than the mean free path mentioned above. When a device is smaller than the phase-relaxation length, the electron waves can interfere. Furthermore, when the cross section of the device becomes as small as the electron wave length (for example, 0.52 nm for gold [33]) the energy levels become quantized, which means that electrons can pass through the system only via specific energy levels. Both these true quantum mechanical phenomena have a large influence on the electronic transport properties.[2]

In the following, we will consider quantum transport only in the form of phase-coherent electrons in nano-scale atomic devices.

3.1.2 The Landauer picture

The simplest model of quantum transport through devices is in terms of single-electron wave functions scattered by a spatially varying potential. One assumes that the potential is situated between two electron reservoirs, each of which emits particles with an equilibrium distribution into the scattering region via perfectly conducting electrodes. The reservoirs will, in general, have different chemical potentials, μ_L and μ_R, their difference $\mu_L - \mu_R$ representing an applied bias voltage. The electrical current conducted by the device corresponds to the net flux of electrons passing between the reservoirs. This picture, which was conceived by Landauer [31], is schematically illustrated in Fig. 3.2.

Let us briefly summarize the formalism that is associated with the Landauer model. In the simplest case, electrons move phase-coherently throughout the device, experiencing only elastic collisions in the scattering region. Any loss of coherence due to inelastic collisions requires a higher-level description.[3] For

[2] For example giving rise to phenomena such as conductance quantization, conspicuous $I - V$ dependences and Coulomb blockade [4].

[3] Meir and Wingreen [34] have, for example, extended the Landauer-Büttiker formalism to incorporate electron-electron interactions in the scattering region.

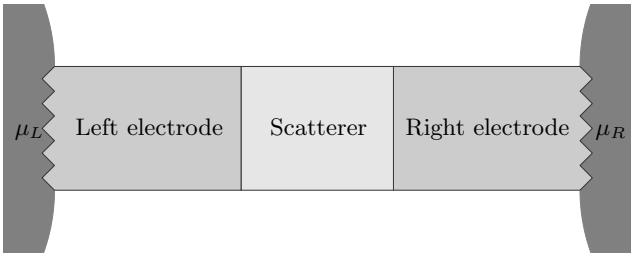


Figure 3.2: Schematic representation of the Landauer picture for electronic transport, where a central scattering region is situated between two electrodes that are connected to thermal reservoirs. What drives the current is the difference between the chemical potentials μ_L and μ_R of the reservoirs.

additional simplicity, we can assume that the electron reservoirs have certain characteristics: (i) the reservoirs are reflectionless, i.e., electrons entering the reservoirs from the electrodes are always accepted, and (ii) the reservoirs are macroscopic in the sense that the chemical potentials are maintained at μ_L and μ_R unaffected by the input and output of electrons via the electrodes. The two assumptions imply that the reservoirs can be described by independent equilibrium Fermi-Dirac distributions, so that the probability that an electron being supplied has energy E is given by Fermi functions

$$f(E-\mu) = \frac{1}{e^{(E-\mu)/k_B \mathrm{T}} + 1}, \tag{3.1}$$

where k_B is Boltzmann's constant, T is the temperature, and μ is the chemical potential of the associated reservoir.

Suppose now that an electron with energy $\mu_R < E < \mu_L$ is provided to the left electrode. The electron is then initially propagating in a right-going Bloch wave in one of the modes of the left electrode. Let us denote the 1D wave number, energy, and group velocity of this mode by k, E_k, and $v_\mathrm{k} = \frac{1}{\hbar}\frac{\mathrm{d}E_\mathrm{k}}{\mathrm{dk}} > 0$, respectively. When the electron hits the central region in the Landauer model, it is exposed to elastic scattering and with some probability T_k transmitted into the right electrode. It will then propagate through the right electrode and finally enter the right reservoir. If it is not transmitted, it is reflected back into the left electrode with probability R_k ($T_\mathrm{k} + R_\mathrm{k} = 1$) and will eventually return to the left reservoir. The current flow of electrons according to such a sequence of events is then given by the first term in the formula

$$I = \frac{e}{L}\sum_{\mathrm{k}} f(E_\mathrm{k} - \mu_L) v_\mathrm{k} T_\mathrm{k} + \frac{e}{L}\sum_{-\mathrm{k}'} f(E_{-\mathrm{k}'} - \mu_R) v_{-\mathrm{k}'} T_{-\mathrm{k}'}, \tag{3.2}$$

where $\frac{e}{L}$ accounts for the electrons being normalized in the 1D volume L. Here the second term corresponds to the reverse flow situation, where the electrons,

designated by wave numbers $-\mathrm{k}'$, are propagating from the right reservoir towards the left ($v_{-\mathrm{k}'} < 0$).

3.1.3 The Landauer-Büttiker formula

To arrive at Eq. (3.2) we considered the transmission probabilities $T_{\pm\mathrm{k}}$ for specific electron wave numbers $\pm\mathrm{k}$. Accordingly, the total transmission coefficients for a particular electron energy E, can be expressed as

$$T^{\pm}(E) = \sum_{\pm k} T_{\pm\mathrm{k}} \delta(E_{\mathrm{k}} - E), \tag{3.3}$$

for the left-going ($-$) and right-going ($+$) electrons, respectively, with δ being the Dirac delta function. Since we must have time reversal symmetry, so that the current is the exact opposite when time is running backwards, the coefficients $T^{-}(E)$ and $T^{+}(E)$ have to be identical: $T^{-}(E) = T^{+}(E) \equiv T(E)$. We can then rewrite Eq. (3.2) as an integral over energies by first transforming the sums over wave numbers into integrals, i.e., $\sum_{\mathrm{k}} \rightarrow 2 \times \frac{L}{2\pi} \int \mathrm{dk}$, where the factor of 2 accounts for spin degeneracy, and subsequently apply $\mathrm{dk} = \mathrm{d}E_{\mathrm{k}}/\hbar v_{\mathrm{k}}$ to obtain

$$I = \frac{2e}{h} \int_{-\infty}^{\infty} T(E)\big[f(E-\mu_L) - f(E-\mu_R)\big]\,\mathrm{d}E. \tag{3.4}$$

which is a key formula in the Landauer-Büttiker theory. We will use it to calculate the current in the numerical examples of this thesis (see Sec. 6.4).

Let us now look at what happens in the zero temperature limit of the above formalism. We note that in this limit the Fermi function in Eq. (3.1) becomes a step function. If we also assume that the difference $|\mu_R - \mu_L|$ is so small that $T(E) \rightarrow T(E_F)$ is independent of E, then Eq. (3.4) can be readily evaluated as

$$I = \frac{2e}{h} T(E_F) \int_{\mu_R}^{\mu_L} \mathrm{d}E = \frac{2e}{h} T(E_F)(\mu_L - \mu_R), \tag{3.5}$$

for the case $\mu_R < \mu_L$. Since the potential difference between the two electrodes is given by $V_b = -(\mu_L - \mu_R)/e$, we arrive at the following expression for the conductance

$$\mathcal{G} = \frac{I}{V_b} = \frac{2e^2}{h} T(E_F), \tag{3.6}$$

which is in agreement with the original Landauer formula from 1957 [31].

Eq. (3.6) implies that the maximum conductance of a single channel with two spin states is $\mathcal{G}_0 = \frac{2e^2}{h}$, for the ballistic case of $T = 1$. The experimental verification of such quantized conductance has been found, e.g., for metal nanowires in break-junction setups [35]. In general, however, the value of quantization of a given channel in a device deviates from $\mathcal{G}_0$, since T is not 1. Instead, we will

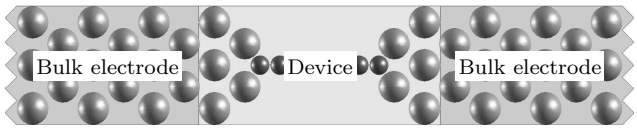


Figure 3.3: Schematic illustration of a nano-scale two-probe system in which a device is sandwiched between two semi-infinite bulk electrodes.

have a total $\mathcal{G}$ that is the sum of contributions of all available channels, with each contribution being less than or equal to $\mathcal{G}_0$.

The conductance is therefore quite intuitively formulated in terms of transmission and reflection matrices, $\mathbf{t}$ and $\mathbf{r}$, that satisfy the unitarity condition $\mathbf{t}^\dagger\mathbf{t} + \mathbf{r}^\dagger\mathbf{r} = \mathbf{1}$, in the case of elastic scattering. The matrix element t_{ij} is the probability amplitude of an incident electron in a mode i in the left electrode being scattered into a mode j in the right electrode, and correspondingly r_{ik} is the probability of it being reflected back into mode k in the left electrode. This simple interpretation then yields the Landauer-Büttiker formula [32]

$$\mathcal{G} = \frac{2e^2}{h}\mathrm{Tr}[\mathbf{t}^\dagger\mathbf{t}], \tag{3.7}$$

which holds in the limit of infinitesimal voltage bias and zero temperature. Consequently, the total transmission for a given electron energy E is

$$T(E) = \mathrm{Tr}[\mathbf{t}^\dagger\mathbf{t}] = \sum_{ij} |t_{ij}|^2, \tag{3.8}$$

where i can be restricted to the Bloch modes of the left electrode and j can be restricted to the Bloch modes of the right electrode. An alternative derivation of Eqs. (3.7) and (3.8) is available from linear response theory [36].

In Secs. 3.2.4 and 3.3.4 we will present numerical procedures to efficiently calculate the transmission $T(E)$ for nano-scale two-probe devices.

3.1.4 Two-probe systems

In correspondence with the Landauer picture, a typical nano-scale device can be conceptually divided into three regions (see Fig. 3.3):

- Left semi-infinite electrode (L) with a periodic bulk configuration
- Central region (C) with an arbitrary device configuration
- Right semi-infinite electrode (R) with a periodic bulk configuration

More precisely, it is clear that far deep in the left and right macroscopic electrodes, the effective potential felt by electrons will resemble that of ideal periodic

bulks because of screening [26]. Thus a central region of appropriate size can always be chosen for which this L–C–R setup is a good approximation. Almost any infinitely extending system with a periodicity breaking region, such as junctions, interfaces, single molecules between electrodes, and tip-sample systems, can be properly modeled in this manner. In the following, we will therefore treat only this type of setup, which we call a two-probe system.

We here start by discussing how to solve the Schrödinger equation for a two-probe system in the LCAO basis of localized non-orthogonal atomic orbitals,

$$(E\mathbf{S} - \mathbf{H})\mathbf{c} = \mathbf{0}, \tag{3.9}$$

where $\mathbf{H}$ is an infinite Hamiltonian matrix with the block form

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_L^{\infty} & \mathbf{H}_{L,C}^{\infty} & \mathbf{0} \\ \mathbf{H}_{L,C}^{\infty\dagger} & \mathbf{H}_C & \mathbf{H}_{R,C}^{\infty} \\ \mathbf{0} & \mathbf{H}_{R,C}^{\infty\dagger} & \mathbf{H}_R^{\infty} \end{pmatrix}, \tag{3.10}$$

and $\mathbf{S}$ is the corresponding overlap matrix with a similar form. The "∞" superscripts indicate blocks of (semi-)infinite order. As such, the key to a proper solution is a practical approach to handle such an infinite eigenvalue problem. This is addressed in detail in Secs. 3.2.1 and 3.3.3. Here, we will first take advantage of the compact support of the LCAO basis in order to obtain a computationally favorable representation of the $\mathbf{H}$ and $\mathbf{S}$ matrices.

Consider for example the Al–C_7–Al two-probe system in Fig. 3.3, where the device configuration corresponds to the central region (C) and the electrodes are two semi-infinite bulks (L and R). We know, that the interaction between distant atoms and the overlap of the corresponding localized orbitals are negligible because of "nearsightedness" [18]. Consequently, there will be sparsity in the Hamiltonian and overlap matrices, where typically only $O(N)$ elements are non-zero in the $N \times N$ central block. There are several ways to exploit this sparsity to end up with a linear scaling method [19].

In this work, we will restrict our efforts to an appropriate layer ordering of the orbitals, which results in computationally attractive block-tridiagonal matrix structures. Let us illustrate the technique with the Al–C_7–Al example for which we have calculated the electronic density $n(\mathbf{r})$ and display the $\hat{\mathbf{x}} \cdot \mathbf{r} = 0$ color coded result in the upper part of Fig. 3.4. We also indicate in this figure how the entire system can be divided into so-called principal layers that only interact with neighbor layers. The orbitals within a given layer are then assumed to be grouped together when indexing the $\mathbf{H}$ and $\mathbf{S}$ matrices. Accordingly, each layer can be described, by appropriate diagonal Hamiltonian matrices $\mathbf{H}_i$ and overlap matrices $\mathbf{S}_i$, where i is the layer number, and off-diagonal matrices written $\mathbf{H}_{i,j}$ and $\mathbf{S}_{i,j}$, that represents the interactions between layers. In this manner the Hamiltonian and overlap matrices becomes block-tridiagonal infinite matrices, which is illustrated in the the bottom part of Fig. 3.4. For the electrodes we

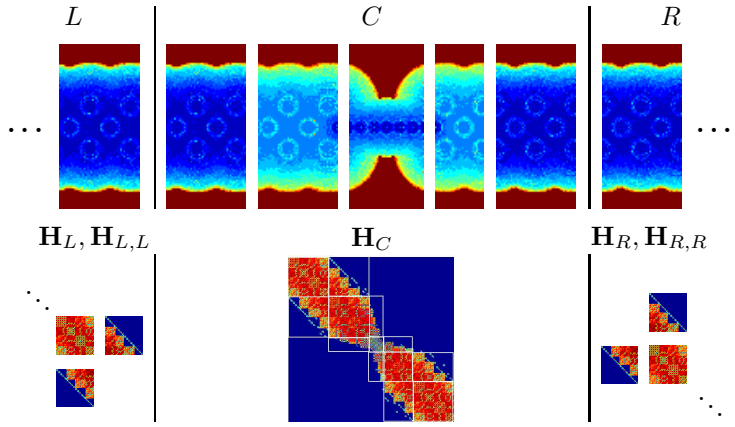


Figure 3.4: Illustration of how the Al–C_7–Al system is modeled by two semi-infinite electrodes (L and R) and a central region (C), each divided into principal layers that only interact with neighbor layers. The upper panel shows the electron density n in the y–z–plane. The lower panel displays the structure of the corresponding Hamiltonian matrices.

use subscripts L and R instead of numbers i, j, since these blocks are repeated throughout the respective electrode.

In the rest of this chapter, we will consequently assume a Hamiltonian, given by

$$\bar{\mathbf{H}} = \begin{pmatrix} \ddots & \ddots & & & & \\ \ddots & \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} & & & \\ & \bar{\mathbf{H}}_{L,L}^\dagger & \begin{pmatrix} & & \\ & \bar{\mathbf{H}}_C & \\ & & \end{pmatrix} & \bar{\mathbf{H}}_{R,R} & & \\ & & & \bar{\mathbf{H}}_{R,R}^\dagger & \bar{\mathbf{H}}_R & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}, \tag{3.11}$$

where the finite matrix of the central device is

$$\bar{\mathbf{H}}_C = \begin{pmatrix} \bar{\mathbf{H}}_1 & \bar{\mathbf{H}}_{1,2} & & \\ \bar{\mathbf{H}}_{1,2}^\dagger & \ddots & \ddots & \\ & \ddots & \ddots & \bar{\mathbf{H}}_{\mathrm{n-1,n}} \\ & & \bar{\mathbf{H}}_{\mathrm{n,n-1}}^\dagger & \bar{\mathbf{H}}_{\mathrm{n}} \end{pmatrix}, \tag{3.12}$$

and where we have introduced notation $\bar{\mathbf{H}} \equiv E\mathbf{S} - \mathbf{H}$. Notice also that the C region in this particular setup contains at least one layer of each electrode,

which means that $\bar{\mathbf{H}}_1 = \bar{\mathbf{H}}_L$ and $\bar{\mathbf{H}}_n = \bar{\mathbf{H}}_R$. This is an important feature of our approach, which we will return to later.

3.1.5 Electronic density

Let us end the brief description of the Landauer-Büttiker theory by linking the derived formalism to the DFT electronic structure method presented in Chap. 2. To do this, the expression for the density $n(\mathbf{r})$ in the Landauer-Büttiker picture is needed. Assuming that the wave functions $\psi_i(\mathbf{r})$ and energies ϵ_i of the KS eigenvalue equations in Eq. (2.4) have been obtained for a given cycle, then, following the same arguments that lead to Eq. (3.2) and Eq. (3.3), we have

$$D(\mathbf{r}, E) = D(\mathbf{r}, E)^+ + D(\mathbf{r}, E)^-, \tag{3.13}$$

$$D(\mathbf{r}, E)^\pm = 2 \sum_{\pm \mathrm{k}} |\psi_{\pm \mathrm{k}}(\mathbf{r})|^2 \delta(E - \epsilon_{\pm \mathrm{k}}), \tag{3.14}$$

for the local density of states (LDOS) in the central region. Again, the +k sum corresponds to right-going (+) electrons incident from the left electrode, and the −k sum to left-going (-) electrons incident from the right. It is here implied that each solution ($\psi_i(\mathbf{r})$,ϵ_i) has a well-defined momentum k_i by which the solutions can be designated $i \rightarrow \mathrm{k}$ in a consistent way.

Integrating the LDOS in the C region over available energies then gives the electronic density in the Landauer-Büttiker approach

$$n(\mathbf{r}) = \int_{-\infty}^{\infty} \left[D(\mathbf{r}, E)^+ f(E - \mu_L) + D(\mathbf{r}, E)^- f(E - \mu_R) \right] \mathrm{d}E, \tag{3.15}$$

where f is again the Fermi function and $\mu_{L/R}$ are the chemical potentials for the left (L) and right (R) reservoirs. It is evident from the distinct terms in Eq. (3.15) that the electronic density in the C region corresponds to two independent contributions, one from electrons coming from the left electrode, in equilibrium with the left reservoir, and one from electrons coming from the right electrode, in equilibrium with the right reservoir. We will come back to the additional steps in the self-consistent DFT procedure which are required to implement the above equations later (see Secs. 3.2.5 and 3.3.5).

3.2 Green's function method

In mathematics, Green's functions are frequently used as a tool to solve differential equations subject to boundary conditions. In the current context, it also turns out to be a convenient approach to solve the infinite Schrödinger equation in Eq. (3.9) for electron transport calculations. The Green's functions have several attractive features that reflects an intuitive interpretation in terms of the electron scattering model at hand. A rigorous treatment of the

electronic properties of nano-scale devices in terms of Green's functions leads to the non-equilibrium Green's function (NEGF) formalism [4, 37, 38]. This is currently the state-of-the-art framework for first principles modeling of the current-voltage relations at the atomic level and much theoretical research has been generated on its success [39, 40, 41, 17, 42, 43, 44, 45, 46, 47].

Since the main focus in this chapter is on the computational aspects of transport calculations, we will here take a numerical point of view on the adoption of the Green's function technique. We leave the details of the properties of Green's functions and their physical interpretation to App. A. The derivations of the primary NEGF formulas from a physical perspective are given in App. B.

3.2.1 Self-energy matrices

To begin with, let us simply reformulate the infinite eigenvalue problem in Eq. (3.9) as a Green's function equation, defined as

$$(E\mathbf{S} - \mathbf{H})\mathbf{G} = \mathbf{I}, \tag{3.16}$$

which for the two-probe system can be written

$$\begin{pmatrix} \bar{\mathbf{H}}_L^{\infty} & \bar{\mathbf{H}}_{L,C}^{\infty} & \mathbf{0} \\ \bar{\mathbf{H}}_{L,C}^{\infty\dagger} & \bar{\mathbf{H}}_C & \bar{\mathbf{H}}_{R,C}^{\infty} \\ \mathbf{0} & \bar{\mathbf{H}}_{R,C}^{\infty\dagger} & \bar{\mathbf{H}}_R^{\infty} \end{pmatrix} \begin{pmatrix} \mathbf{G}_L^{\infty} & \mathbf{G}_{L,C}^{\infty} & \mathbf{G}_{L,R}^{\infty} \\ \mathbf{G}_{L,C}^{\infty\dagger} & \mathbf{G}_C & \mathbf{G}_{R,C}^{\infty} \\ \mathbf{G}_{L,R}^{\infty\dagger} & \mathbf{G}_{R,C}^{\infty\dagger} & \mathbf{G}_R^{\infty} \end{pmatrix} = \begin{pmatrix} \mathbf{I}^{\infty} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}^{\infty} \end{pmatrix}, \tag{3.17}$$

and set out to solve for the Green's function matrix $\mathbf{G}$ instead of the wave function $\mathbf{c}$. We have again used the notation $\bar{\mathbf{H}} \equiv E\mathbf{S} - \mathbf{H}$ with the significant difference that an imaginary part is added to the energy, i.e., $E \to E + \hat{\imath}\eta$, where η is an infinitesimal quantity. This is to specify the proper retarded Green's function from the two available solutions, as explained in App. A.

The motivation for transforming Eq. (3.9) into Eq. (3.16) from a numerical perspective, is the interest in obtaining the solution of the Schrödinger equation only for the central C part of the two-probe system, while being able to accommodate the correct influence of the electrode parts. Since the C region is where all the scattering occurs, it must include the full information of the electronic transport processes, in the sense of the Landauer picture in Sec. 3.1.2 [4].

Consider the linear system in Eq. (3.17) and, for the moment, ignore that its size is infinite. We then see that, by subtraction the appropriate multiplication of the top and bottom block rows from the central block row, it is possible to eliminate the blocks $\bar{\mathbf{H}}_{L,C}^{\infty\dagger}$ and $\bar{\mathbf{H}}_{R,C}^{\infty}$ from the leftmost matrix. To be more specific, we perform a block Gaussian elimination by which the blocks mentioned vanish and central block is modified accordingly $\bar{\mathbf{H}}_C \to \bar{\mathbf{H}}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R$, where

$$\mathbf{\Sigma}_L = \bar{\mathbf{H}}_{L,C}^{\infty\dagger}(\bar{\mathbf{H}}_L^{\infty})^{-1}\bar{\mathbf{H}}_{L,C}^{\infty}, \tag{3.18}$$

and

$$\mathbf{\Sigma}_R = \bar{\mathbf{H}}_{R,C}^{\infty}(\bar{\mathbf{H}}_R^{\infty})^{-1}\bar{\mathbf{H}}_{R,C}^{\infty\dagger}. \tag{3.19}$$

The expressions in Eqs. (3.18) and (3.19) are simply the second terms of the Schur's complements from the top and bottom elimination operations. In this manner, we end up with a central block row $(\mathbf{0}, \bar{\mathbf{H}}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R, \mathbf{0})^T$ to be multiplied onto the central column of $\mathbf{G}$ and equal to $\mathbf{I}$, which isolates an expression for $\mathbf{G}_C$, given by

$$\mathbf{G}_C = ((E + \hat{\imath}\eta)\mathbf{S}_C - \mathbf{H}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R)^{-1}, \tag{3.20}$$

where the notation for $\bar{\mathbf{H}}_C$ is explicitly written out.

It might seem that we have not gained much from deriving Eqs. (3.18)–(3.20) since the formulas to obtain $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ still involve semi-infinite matrices. In practice, however, because the layered two-probe setup generates Hamiltionian and overlap matrices with block-tridiagonal structure, the elements of $\bar{\mathbf{H}}^{\infty}_{L,C}$ are all zero except in the lower left corner, where they are equal to $\bar{\mathbf{H}}_{L,L}$ of size $m_L \times m_L$. Likewise, for $\bar{\mathbf{H}}^{\infty\dagger}_{L,C}$, $\bar{\mathbf{H}}^{\infty}_{R,C}$, and $\bar{\mathbf{H}}^{\infty\dagger}_{R,C}$, which are non-zero only in the blocks $\bar{\mathbf{H}}^{\dagger}_{L,L}$,$\bar{\mathbf{H}}_{R,R}$, and $\bar{\mathbf{H}}^{\dagger}_{R,R}$, respectively (see Eq. (3.11)). This allows us to determine the finite number of non-zero elements of the $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ from expressions with finite matrices, i.e.,

$$[\mathbf{\Sigma}_L]_{m_L \times m_L} = \bar{\mathbf{H}}^{\dagger}_{L,L}\mathbf{g}_L\bar{\mathbf{H}}_{L,L}, \tag{3.21}$$

and

$$[\mathbf{\Sigma}_R]_{m_R \times m_R} = \bar{\mathbf{H}}_{R,R}\mathbf{g}_R\bar{\mathbf{H}}^{\dagger}_{R,R}, \tag{3.22}$$

where $\mathbf{g}_L$ and $\mathbf{g}_R$ are the so-called surface Green's functions of the electrodes corresponding to the lower right $m_L \times m_L$ submatrix of $\mathbf{G}^{\infty}_L \equiv (\bar{\mathbf{H}}^{\infty}_L)^{-1}$ and the upper left $m_R \times m_R$ submatrix of $\mathbf{G}^{\infty}_R \equiv (\bar{\mathbf{H}}^{\infty}_R)^{-1}$, respectively.

Finally, by using Eq. (3.12), we arrive at the following matrix inversion operation in order to obtain the central part of the Green's function

$$\mathbf{G}_C = \begin{pmatrix} \bar{\mathbf{H}}_1 - \mathbf{\Sigma}_L & \bar{\mathbf{H}}_{1,2} & & \\ \bar{\mathbf{H}}^{\dagger}_{1,2} & \ddots & \ddots & \\ & \ddots & \ddots & \bar{\mathbf{H}}_{\mathrm{n}-1,\mathrm{n}} \\ & & \bar{\mathbf{H}}^{\dagger}_{\mathrm{n}-1,\mathrm{n}} & \bar{\mathbf{H}}_{\mathrm{n}} - \mathbf{\Sigma}_R \end{pmatrix}^{-1}, \tag{3.23}$$

where the $[\cdot]_{m\times m}$ notation is implied. From here on, we will call the matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ the self-energy matrices of the left (L) and right (R) electrodes [4].

3.2.2 Surface Green's functions

So far, the main computational tasks identified in the Green's function method are the evaluation of the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$, and the subsequent calculation of $\mathbf{G}_C$ by means of matrix inversion. As will become clear during this chapter, these tasks are actually the foundation for most other necessary

calculations, including the computation of the density matrix and current. In fact, for almost all systems, the most overall CPU time is spent (see Sec. 3.5.1) determining the surface Green's function $\mathbf{g}_L$ and $\mathbf{g}_R$ that are needed for the self-energy matrices via Eqs. (3.21)–(3.22). It is therefore relevant to discuss the details of this very costly step of the Green's function approach.

As introduced above, the surface Green's function $\mathbf{g}_L$ depends only on the semi-infinite left part of the two-probe Hamiltonian in Eq. (3.11) and the energy E (including the infinitesimal η). Consider the matrix equation $(E\mathbf{S}_L^\infty - \mathbf{H}_L^\infty)\mathbf{G}_L^\infty = \bar{\mathbf{H}}_L^\infty \mathbf{G}_L^\infty = \mathbf{I}^\infty$ that defines $\mathbf{G}_L^\infty$, written as

$$\begin{pmatrix} \ddots & \ddots & \\ \ddots & \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} \\ & \bar{\mathbf{H}}_{L,L}^\dagger & \bar{\mathbf{H}}_L \end{pmatrix} \begin{pmatrix} \ddots & \vdots & \vdots & \vdots \\ \cdots & \mathbf{G}_L^{n-2,n-2} & \mathbf{G}_L^{n-2,n-1} & \mathbf{G}_L^{n-2,n} \\ \cdots & \mathbf{G}_L^{n-1,n-2} & \mathbf{G}_L^{n-1,n-1} & \mathbf{G}_L^{n-1,n} \\ \cdots & \mathbf{G}_L^{n,n-2} & \mathbf{G}_L^{n,n-1} & \mathbf{G}_L^{n,n} \end{pmatrix} = \mathbf{I}^\infty, \tag{3.24}$$

where the new block notation $\mathbf{A}^{i,j} = \mathbf{A}_{i,j} \equiv [\mathbf{A}]_{ij}$ of a matrix $\mathbf{A}$ was used. It seems that in order to obtain the surface Green's function matrix $\mathbf{g}_L = \mathbf{G}_L^{n,n}$ from Eq. (3.24), we apparently have to perform an infinite number of block Gaussian eliminations. Fortunately, such a succession of eliminations in the case of the block-Toeplitz matrix $\bar{\mathbf{H}}_L^\infty$ will often converge, in the sense that the Schur complements of two consecutive eliminations become identical to machine precision, after a finite number of iterations.

Several well-established methods exist for this [48]. For comparison, we here discuss three different schemes to find the surface Green's functions of the semi-infinite electrodes and also provide implementation details.

Recursive method

It is straightforward to obtain a recursive expression for $\mathbf{g}_L$ by considering the above Hamiltonian $\bar{\mathbf{H}}_L^\infty$ of the semi-infinite left electrode and a similar Hamiltonian $\bar{\mathbf{H}}_L^{\infty+1}$ with one extra principal layer on the surface, i.e.,

$$\bar{\mathbf{H}}_L^\infty = \begin{pmatrix} \ddots & \ddots & \\ \ddots & \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} \\ & \bar{\mathbf{H}}_{L,L}^\dagger & \bar{\mathbf{H}}_L \end{pmatrix}, \quad \bar{\mathbf{H}}_L^{\infty+1} = \begin{pmatrix} \begin{pmatrix} & \bar{\mathbf{H}}_L^\infty & \\ & & \bar{\mathbf{H}}_{L,L}^\dagger \end{pmatrix} & \begin{matrix} \bar{\mathbf{H}}_{L,L} \\ \bar{\mathbf{H}}_L \end{matrix} \end{pmatrix}. \tag{3.25}$$

Since the matrix $\bar{\mathbf{H}}_L^{\infty+1}$ has an explicit 2×2 block structure and $\mathbf{g}_L^{\infty+1} = [(\bar{\mathbf{H}}_L^{\infty+1})^{-1}]_{2,2}$ we simply apply the matrix inverse identity in Eq. (1.2) to write

$$\mathbf{g}_L^{(n+1)} = [\bar{\mathbf{H}}_L - \bar{\mathbf{H}}_{L,L}^\dagger \mathbf{g}_L^{(n)} \bar{\mathbf{H}}_{L,L}]^{-1}, \tag{3.26}$$

where it is used that $\mathbf{G}_L^\infty = (\bar{\mathbf{H}}_L^\infty)^{-1}$. Here ∞ has been replaced by (n) to indicate a finite iteration number. A similar derivation for the right electrode

gives the expression

$$\mathbf{g}_R^{(n+1)} = [\bar{\mathbf{H}}_R - \bar{\mathbf{H}}_{R,R}\mathbf{g}_R^{(n)}\bar{\mathbf{H}}_{R,R}^{\dagger}]^{-1}, \tag{3.27}$$

for the evaluation of $\mathbf{g}_R^{\infty+1} \equiv [(\bar{\mathbf{H}}_R^{\infty+1})^{-1}]_{1,1}$. Notice that the formulas for $\mathbf{g}_L$ and $\mathbf{g}_R$ differ only in the order of the off-diagonal blocks in the second term because the former represents downwards block Gaussian elimination and the latter upwards block Gaussian elimination.

Eqs. (3.26) and (3.27) form the basis for simple recursive algorithms that use $\mathbf{g}_L^{(0)} = \bar{\mathbf{H}}_L^{-1}$ and $\mathbf{g}_R^{(0)} = \bar{\mathbf{H}}_R^{-1}$ as initialization and iterate until the change from adding an extra principal layer is less than a given tolerance. The fastest way to implement this is by avoiding the explicit inverse, for example

ALGORITHM I: *Recursive method*

1. initialize $\mathbf{A} := \bar{\mathbf{H}}_L$
2. solve $\mathbf{AX} = \bar{\mathbf{H}}_{L,L}$
3. $\mathbf{A} := \bar{\mathbf{H}}_L - \bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{X}$

   iterate until $(\mathbf{A}' - \mathbf{A})_{ij}^2 \leq \delta^2$,
4. $\mathbf{g}_L := \mathbf{A}^{-1}$

(3.28)

for the left electrode, resulting in $\mathbf{g}_L$ of accuracy δ at convergence. It is clear, however, in the light of the available method described next, that this recursive method is not of practical interest On the other hand it is easy to implement as a reference for other methods.

Recursive 2^n method

The most commonly used method in practice is probably the recursive technique of Lopez-Sancho *et. al.* [49], which has exponential convergence in the number of iterations. We can derive this procedure by looking at the last column of the left electrode Green's function $\mathbf{G}_L^{\infty}$ in Eq. (3.24), which can be written

$$\begin{aligned} \bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{G}_L^{n-1,n} + \bar{\mathbf{H}}_L\mathbf{G}_L^{n,n} &= \mathbf{I}, \\ \bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{G}_L^{i-1,n} + \bar{\mathbf{H}}_L\mathbf{G}_L^{i,n} + \bar{\mathbf{H}}_{L,L}\mathbf{G}_L^{i+1,n} &= \mathbf{0}, \quad (i < n). \end{aligned} \tag{3.29}$$

Consider three succeeding equations, e.g., for $i-1$, i and $i+1$, from this chain of mutually dependent linear equations. We can use the first and last of these equations to isolate expressions for $\mathbf{G}_L^{i-1,n}$ and $\mathbf{G}_L^{i+1,n}$, given by

$$\begin{aligned} \mathbf{G}_L^{i-1,n} &= -\bar{\mathbf{H}}_L^{-1}[\bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{G}_L^{i-2,n} + \bar{\mathbf{H}}_{L,L}\mathbf{G}_L^{i,n}], \\ \mathbf{G}_L^{i+1,n} &= -\bar{\mathbf{H}}_L^{-1}[\bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{G}_L^{i,n} + \bar{\mathbf{H}}_{L,L}\mathbf{G}_L^{i+2,n}], \end{aligned} \tag{3.30}$$

and subsequently insert them into the equation for i. After rearranging the terms, this leads to the following equation

$$\bar{\mathbf{H}}_{L,L}^{\dagger}\bar{\mathbf{H}}_{L}^{-1}\bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{G}_{L}^{i-2,n} + [\bar{\mathbf{H}}_{L} - \bar{\mathbf{H}}_{L,L}^{\dagger}\bar{\mathbf{H}}_{L}^{-1}\bar{\mathbf{H}}_{L,L} - \bar{\mathbf{H}}_{L,L}\bar{\mathbf{H}}_{L}^{-1}\bar{\mathbf{H}}_{L,L}^{\dagger}]\mathbf{G}_{L}^{i,n} + \bar{\mathbf{H}}_{L,L}\bar{\mathbf{H}}_{L}^{-1}\bar{\mathbf{H}}_{L,L}\mathbf{G}_{L}^{i+2,n} = \mathbf{0} \quad (3.31)$$

which can be written in the simple form

$$\mathbf{A}_{22}\mathbf{G}_{L}^{i-2,n} + \mathbf{B}\mathbf{G}_{L}^{i,n} + \mathbf{A}_{11}\mathbf{G}_{L}^{i+2,n} = \mathbf{0}, \quad (i < n-2), \quad (3.32)$$

where we have defined two new matrices $\mathbf{A}$ and $\mathbf{B}$, given by

$$\mathbf{A} = \begin{pmatrix} \bar{\mathbf{H}}_{L,L} \\ \bar{\mathbf{H}}_{L,L}^{\dagger} \end{pmatrix} \bar{\mathbf{H}}_{L}^{-1} \begin{pmatrix} \bar{\mathbf{H}}_{L,L} & \bar{\mathbf{H}}_{L,L}^{\dagger} \end{pmatrix}, \quad \mathbf{B} = \bar{\mathbf{H}}_{L} - \mathbf{A}_{21} - \mathbf{A}_{12}. \quad (3.33)$$

Moreover, if we use $i = n$ in the upper equation of Eq. (3.30) and insert this in the initial expression of Eq. (3.29), we get

$$\mathbf{A}_{22}\mathbf{G}_{L}^{n-2,n} + \mathbf{C}\mathbf{G}_{L}^{n,n} = \mathbf{I}, \quad (3.34)$$

where $\mathbf{C} = \bar{\mathbf{H}}_{L} - \mathbf{A}_{21}$ has been introduced. Now notice that equation Eq. (3.32), which is valid for all $i < n-2$, together with equation Eq. (3.34), has the same form as the original chain of equations in Eq. (3.29), if we only include even values of i, i.e. $i \to 2i$ in Eq. (3.32). This means that the original and new chain can be classified as isomorphic, where the latter obviously has twice the spacing between principal layers of the former.

The final step is to realize that there is nothing that prevents us from applying the above replacements again on the new chain of equations, and thus repeatedly. Then the result is another recursive scheme for the evaluation of $\mathbf{g}_L = \mathbf{G}_L^{n,n}$, that with each iteration doubles the number of principal layers taken into account, that is 2^n layers for n iterations. We will implement the method in the following simple way:

ALGORITHM II: *Recursive* 2^n *method*

1. initialize $\mathbf{A} := \begin{pmatrix} \mathbf{H}_{LL} & 0 \\ 0 & \mathbf{H}_{LL}^{\dagger} \end{pmatrix}$, $\mathbf{B} := \mathbf{H}_L$, $\mathbf{C} := \mathbf{H}_L$
2. $\mathbf{A} := \begin{pmatrix} \mathbf{A}_{11} \\ \mathbf{A}_{22} \end{pmatrix} \mathbf{B}^{-1} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{22} \end{pmatrix}$
3. $\mathbf{B} := \mathbf{B} - \mathbf{A}_{12} - \mathbf{A}_{21}$
4. $\mathbf{C} := \mathbf{C} - \mathbf{A}_{12}$

   iterate until $(\mathbf{A}_{12})_{ij}^2 \leq \delta^2$
5. $\mathbf{g}_L := \mathbf{C}^{-1}$

(3.35)

where δ sets the accuracy (we use machine precision 10^{-15}). Here $\mathbf{A}$ can be calculated without finding the explicit inverse $\boldsymbol{B}^{-1}$ in the same manner as for

Eq. (3.28). After convergence the surface Green's function of the left electrode can be obtained as $\mathbf{g}_L = \mathbf{C}^{-1}$. In the case of a semi-infinite right electrode the last line of the iteration step 4 would simply be $\mathbf{C} := \mathbf{C} - \mathbf{A}_{21}$, resulting in $\mathbf{g}_R = \mathbf{C}^{-1}$ at completion.

Notice that n iterations of the above algorithm corresponds exactly to $2^n - 1$ iterations of the recursive formula Eq. (3.28). This technique is therefore much faster and more stable in terms of avoiding rounding errors than the linear recursive scheme described previously [48].

Other methods

We will briefly mention a few other schemes which have been developed for calculating surface Green's functions in order to obtain self-energy matrices.

First we note that in the very limited number of cases where the off-diagonal matrices $\bar{\mathbf{H}}_{L,L}$ and $\bar{\mathbf{H}}_{R,R}$ are non-singular, the surface Green's functions can be calculated in closed form, e.g., as devised by Umerski [50]. Such an approach is somewhat faster even than the recursive 2^n method, but obviously limited in applicability and therefore rarely used.

Alternatively, Rocha and Sanvito *et. al.* [51] attacks the problem of singular off-diagonal matrices directly with a regularization approach that transforms the full block tridiagonal Schrödinger equation into an equivalent block tridiagonal equation with well-conditioned coupling matrices. The surface Green's functions are then determined from the Bloch solutions of the transformed electrode Hamiltonians in a manner that is both fast and stable. In this approach they argue that the regularization is a necessary step in order to reduce errors during the process of constructing $\mathbf{g}_L$ and $\mathbf{g}_R$.

On the other hand, the evaluation of the surface Green's functions from the Bloch solutions can be attempted without any prior regularization (and with real E and $\eta = 0$). The key is to omit the solutions that are unphysical due to the mentioned singularities. We credit the first application of this approach to Ando [52] and adopt it here in connection with the wave function matching method described later in this chapter. Over the years the method has been modified and refined by other authors [53, 54, 55]. We believe that our implementation for obtaining the electrode self-energy matrices using this approach (see Secs. 3.3.2 and 3.3.3) avoids the stability problems mentioned by Rocha and Sanvito, and therefore renders an explicit regularization superfluous. We also note that in the most recent and still unpublished work [56] from the group of Sanvito, the authors also consider situations where the regularization is not needed, which is fortunately always the case in this work.

3.2.3 Matrix inversion

We now turn to the second computationally heavy and repeatedly executed task in the Green's function method, which is the matrix inversion operation

to obtain $\mathbf{G}_C$. In principle, although the Hamiltonian matrix to be inverted in Eq. (3.23) has block tridiagonal structure, the inverse matrix is going to be filled out. The inverse operation is then at least $O(N^2)$ in complexity even when $\bar{\mathbf{H}}_C - \Sigma_L - \Sigma_R$ has $O(N)$ elements. We remind the reader that $N \approx \mathrm{n}\bar{m}$, where n is the number of blocks in $\bar{\mathbf{H}}_C$ and $\bar{m}$ is the maximum size of any of the blocks (i.e., $O(\mathrm{n}\bar{m}^2) \sim O(N)$ elements). Fortunately, as we will see in Sec. 3.2.5, the main quantities needed in the Green's function method can be obtained without knowing the entire $\mathbf{G}_C$ matrix. In particular, the density matrix $\mathbf{n}$ requires only the familiar block tridiagonal part of $\mathbf{G}_C$, which corresponds to sites where the localized orbitals overlap. This part has only $O(N)$ elements and can be obtained in $O(N)$ complexity. We will show how to do this in Chap. 4, but first look at a standard full matrix inverse procedure.

Consider the block Gaussian elimination method introduced in Eqs. (1.3)–(1.6). The particular downwards "sweep" performed has eliminated the lower off-diagonal blocks. Let us attempt to complete the transformation of the left-hand side of the augmented matrix $(\mathbf{A}|\mathbf{I})$ in Eq. (1.4) until it becomes the identity matrix. We know that when this is the case, the resulting right-hand side matrix will be the inverse matrix. Therefore, we will do an upwards block Gaussian elimination "sweep" to eliminate the upper off-diagonal blocks in Eq. (1.4)

$$\text{Eq. (1.4)} \sim \left(\begin{array}{cccc|cccc} \mathbf{A}'_{11} & & & & \mathbf{J}'_{11} & \mathbf{J}'_{1,2} & \cdots & \mathbf{J}'_{1,\mathrm{n}} \\ & \mathbf{A}'_{22} & & & \mathbf{J}'_{2,1} & \ddots & \ddots & \vdots \\ & & \ddots & & \vdots & \ddots & \mathbf{J}'_{\mathrm{n}-1,\mathrm{n}-1} & \mathbf{J}'_{\mathrm{n}-1,\mathrm{n}} \\ & & & \mathbf{A}'_{\mathrm{n,n}} & \mathbf{J}'_{\mathrm{n},1} & \cdots & \mathbf{J}'_{\mathrm{n},\mathrm{n}-1} & \mathbf{I} \end{array}\right), \tag{3.36}$$

where $\mathbf{A}'_{ii}$ and $\mathbf{J}'_{\mathrm{n},i} = \mathbf{J}_{\mathrm{n},i}$ are defined in Eqs. (1.5)–(1.6), and

$$\mathbf{J}'_{ij} = \mathbf{J}_{ij} - \mathbf{A}_{i,i+1}(\mathbf{A}'_{i+1,i+1})^{-1}\mathbf{J}'_{i+1,j}, \quad 1 \le i < \mathrm{n}, 1 \le j < i, \tag{3.37}$$

$$\mathbf{J}'_{ii} = \mathbf{I} - \mathbf{A}_{i,i+1}(\mathbf{A}'_{i+1,i+1})^{-1}\mathbf{J}'_{i+1,i}, \quad 1 \le i < \mathrm{n}. \tag{3.38}$$

$$\mathbf{J}'_{ij} = -\mathbf{A}_{i,i+1}(\mathbf{A}'_{i+1,i+1})^{-1}\mathbf{J}'_{i+1,j}, \quad 1 \le i < \mathrm{n}, i < j \le \mathrm{n}, \tag{3.39}$$

Now we can simply LU-factorize the diagonal matrices $\mathbf{A}'_{ii}$ in Eq. (3.36) and multiply $(\mathbf{A}'_{ii})^{-1}$ onto the ith row of the augmented matrix, which yields the identity matrix in the left-hand side and $\mathbf{A}^{-1}$ in the right-hand side. Explicitly counting the basic block operations of this upwards sweep procedure results in n LU-factorizations, $2\mathrm{n}^2 + \mathrm{n} - 2$ multiplications and $\frac{1}{2}(\mathrm{n}^2 + \mathrm{n} - 2)$ additions, which means that it is of $O(\mathrm{n}^2\bar{m}^3) \sim O(N^2)$ complexity, as expected.

Accordingly, the basic algorithm that can be implemented in order to calculate the Green's function $\mathbf{G}_C$ of the central region for two-probe systems is

written:

ALGORITHM III: *Green's function method*

1. obtain $\mathbf{g}_L$ using ALGORITHM II
2. $\mathbf{\Sigma}_L := \bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{g}_L\bar{\mathbf{H}}_{L,L}$

} the left elecrode

3. obtain $\mathbf{g}_R$ using ALGORITHM II
4. $\mathbf{\Sigma}_R := \bar{\mathbf{H}}_{R,R}\mathbf{g}_R\bar{\mathbf{H}}_{R,R}^{\dagger}$

} the right elecrode

5. initialize $\mathbf{A}'_1 := \bar{\mathbf{H}}_{11} - \mathbf{\Sigma}_L$
6. **for** $i := 2, \ldots, \mathrm{n}$
7. solve $\bar{\mathbf{H}}_{i,i-1} = \mathbf{J}_{i,i-1}\mathbf{A}'_{i-1}$ for $\mathbf{J}_{i,i-1}$
8. $\mathbf{A}'_i := \bar{\mathbf{H}}_i - \mathbf{J}_{i,i-1}\bar{\mathbf{H}}_{i-1,i}$
9. **for** $j := 2, \ldots, i-1$
10. $\mathbf{J}_{i,j} := -\mathbf{J}_{i,i-1}\mathbf{J}_{i-1,j}$
11. **end**
12. **end**

} downwards sweep

13. initialize $\mathbf{A}'_{\mathrm{n}} := \mathbf{A}'_{\mathrm{n}} - \mathbf{\Sigma}_R,\ \mathbf{J}'_{\mathrm{n},j} = \mathbf{J}_{\mathrm{n},j}$ for $j = 1, \ldots, \mathrm{n}$ (3.40)
14. **for** $i := \mathrm{n}-1, \ldots, 1$
15. solve $\bar{\mathbf{H}}_{i,i+1} = \mathbf{J}'_{i,i+1}\mathbf{A}'_{i+1}$ for $\mathbf{J}'_{i,i+1}$
16. **for** $j := i+2, \ldots, \mathrm{n}$
17. $\mathbf{J}'_{i,j} := -\mathbf{J}'_{i,i+1}\mathbf{J}'_{i+1,j}$
18. **end**
19. $\mathbf{J}'_{i,i} := \mathbf{I} - \mathbf{J}'_{i,i+1}\mathbf{J}'_{i+1,j}$
20. **for** $j := 1, \ldots, i-1$
21. $\mathbf{J}'_{i,j} := \mathbf{J}'_{i,j} - \mathbf{J}'_{i,i+1}\mathbf{J}'_{i+1,j}$
22. **end**
23. **end**

} upwards sweep

24. **for** $i := 1, \ldots, \mathrm{n}$
25. solve $\mathbf{A}'_i[\mathbf{G}_{i,1}\ \mathbf{G}_{i,2}\ \cdots\ \mathbf{G}_{i,\mathrm{n}}] = [\mathbf{J}'_{i,1}\ \mathbf{J}'_{i,2}\ \cdots\ \mathbf{J}'_{i,\mathrm{n}}]$
26. **end**

} $\mathbf{G}_C$

The total number of basic block operations for this Green's function algorithm is $3\mathrm{n}-2$ LU-factorizations, $\mathrm{n}^2+4\mathrm{n}-4$ multiplications and $4\mathrm{n}-6$ additions.

3.2.4 Transmission calculations

The primary objective in the current modeling of quantum transport is to calculate the current-voltage characteristics of specific nano-scale systems. In order

to do this using the Green's function method we will apply the NEGF formula for the current I through a two-probe system in the coherent limit, which is properly derived in App. B. The formula for a small bias $V_b = -(\mu_L - \mu_R)/e$ is written as

$$I = \frac{2e}{h}\int_{-\infty}^{\infty} T(E)\left(f(E-\mu_L) - f(E-\mu_R)\right)\,\mathrm{d}E, \tag{3.41}$$

where f is the Fermi function in Eq. (3.1), and

$$T(E) = \mathrm{Tr}\{\mathbf{\Gamma}_L \mathbf{G}_C^\dagger \mathbf{\Gamma}_R \mathbf{G}_C\}, \tag{3.42}$$

is the transmission at energy E. Here we have introduced the matrices

$$\mathbf{\Gamma}_L = i(\mathbf{\Sigma}_L - \mathbf{\Sigma}_L^\dagger), \quad \mathbf{\Gamma}_R = i(\mathbf{\Sigma}_R - \mathbf{\Sigma}_R^\dagger), \tag{3.43}$$

which we will call the broadening matrices. The expression for the transmission coefficient in Eq. (3.42) is often referred to as the Caroli formula [57]. In practice, the limits of the integral in Eq. (3.41) can be restricted to a finite range because $f(E-\mu)$ goes rapidly towards zero when E differs from μ, as described in the previous section. We will now focus on the computational technique used to calculate the transmission $T(E)$ at a given energy E from Eq. (3.42).

It is evident that all the quantities in the equations above can be determined from the matrices $\mathbf{\Sigma}_L$,$\mathbf{\Sigma}_R$, and $\mathbf{G}_C$ introduced previously. Since we have already discussed how to obtain $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ in an efficient manner, and subsequently $\mathbf{G}_C$ by matrix inversion, all requisites are available to use Eq. (3.42). However, we actually only need a single block of $\mathbf{G}_C$ to find $T(E)$ because the broadening matrices $\mathbf{\Gamma}_L$ and $\mathbf{\Gamma}_R$ exists only in the corners blocks of the central region matrix structure like $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$. More specifically, we have the simplification

$$\begin{aligned} T(E) &= \mathrm{Tr}\left\{ \begin{pmatrix} \mathbf{\Gamma}_L & \mathbf{0} & & \\ \mathbf{0} & \ddots & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ & & \mathbf{0} & \mathbf{0} \end{pmatrix} \times \mathbf{G}_C^\dagger \times \begin{pmatrix} \mathbf{0} & \mathbf{0} & & \\ \mathbf{0} & \ddots & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ & & \mathbf{0} & \mathbf{\Gamma}_R \end{pmatrix} \times \mathbf{G}_C \right\}, \\ &= \mathrm{Tr}\{\mathbf{\Gamma}_L \mathbf{G}_{\mathrm{n},1}^\dagger \mathbf{\Gamma}_R \mathbf{G}_{\mathrm{n},1}\}, \end{aligned} \tag{3.44}$$

where $\mathbf{G}_{\mathrm{n},1} \equiv [\mathbf{G}_C]_{\mathrm{n},1}$, which implies that only the upper right block of the inverse of $\mathbf{H}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R$ is necessary for the evaluation of $T(E)$.

From Eq. (3.44) we see explicitly why modeling quantum transport in terms of Green's functions is convenient. The information required to find the transmission between layer 1 and layer n is fully contained in the Green's function block $\mathbf{G}_{\mathrm{n},1}$ because the Green's function does not represent the single-particle wave functions themselves, but rather the wave functions resulting from a unit excitation somewhere in the system. In particular, the block $\mathbf{G}_{i,j}$ describes the wave functions in layer i resulting from the possible excitations of orbitals local

to layer j, so that $\mathbf{G}_{\mathrm{n},1}$ gives directly the probabilities of observing electrons in the right electrode as a result of incident electrons from the left electrode.

Again, it is relatively simple to implement an algorithm for computing $\mathbf{G}_{\mathrm{n},1}$ using block Gaussian eliminations. Considering the downwards block Gaussian elimination defined in Eqs. (1.3)–(1.4), we see that such a single sweep is sufficient to obtain the inverse of the entire bottom row of the block tridiagonal matrix, and hence also the block $[(\bar{\mathbf{H}}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R)^{-1}]_{\mathrm{n},1}$. The following sequential algorithm can therefore be implemented to obtain $T(E)$ at completion:

ALGORITHM IV: *Calculate* $T(E)$ *using* $\mathbf{G}_{\mathrm{n},1}$

1. initialize $\mathbf{A} := \bar{\mathbf{H}}_{1,1} - \mathbf{\Sigma}_L,\ \mathbf{B} := \mathbf{I}$
2. **for** $i := 2, \ldots, \mathrm{n}$
3. solve $\bar{\mathbf{H}}_{i,i-1} = \mathbf{X}\mathbf{A}$ for $\mathbf{X}$
4. $\mathbf{A} := \bar{\mathbf{H}}_{i,i} - \mathbf{X}\bar{\mathbf{H}}_{i-1,i}$ (3.45)
5. $\mathbf{B} := -\mathbf{X}\mathbf{B}$
6. **end**
7. solve $(\mathbf{A} - \mathbf{\Sigma}_L)\mathbf{G}_{\mathrm{n},1} = \mathbf{B}$
8. obtain $T(E)$ from Eqs. (3.43) and (3.44),

We note that this algorithm has an $O(\mathrm{n}\bar{m}^3)$ computational complexity, where $\bar{m}$ is the maximum order of any block in $\bar{\mathbf{H}}_C$ (i.e., it is $O(N)$ in the total number of orbitals N, since $N \approx \mathrm{n}\bar{m}$).

We will take advantage of the useful physical interpretation of $\mathbf{G}_{i,j}$ mentioned above to develop a new algorithm in Chap. 4, which is generally faster than the algorithm in Eq. (3.45). A nicely performing parallel algorithm for obtaining $\mathbf{G}_{\mathrm{n},1}$ has been published in [58].

3.2.5 Obtaining the Self-consistent Hamiltonian

As introduced in Chap. 2, we will apply DFT in order to obtain the Hamiltonian matrices $\mathbf{H}_L$,$\mathbf{H}_{L,L}$,$\mathbf{H}_R$,$\mathbf{H}_{R,R}$, and $\mathbf{H}_C$ for the Landauer-Büttiker type two-probe systems, which we are concerned with here. As we have argued several times above, the assumption that makes this computationally feasible for an open and infinite system, is that we are able to separate it into three distinct regions: bulk–[central region]–bulk. Consequently, this involves solving the Schrödinger equation and the Poisson equation in a self-consistently manner for each region of the two-probe system separately.

Consider the self-consistent DFT procedure in the flow chart in Fig. 2.1. To begin with, the self-consistent electrode Hamiltonians $\mathbf{H}_L$ and $\mathbf{H}_R$ are calculated according to this procedure by properly setting them up as ideal periodic systems using **k**-point sampling (see Sec. 2.1.7). Also the off-diagonal matrices $\mathbf{H}_{L,L}$ and $\mathbf{H}_{R,R}$ can be constructed from the self-consistent solution. The unit cells

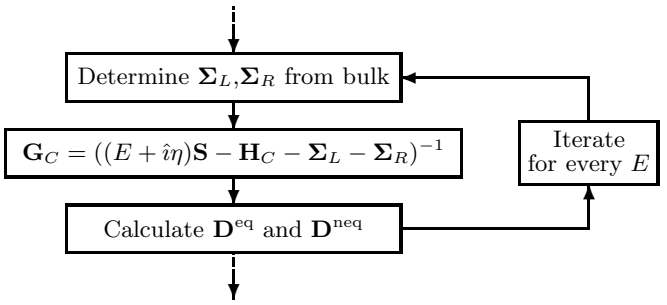


Figure 3.5: Detailed flow-diagram for the "solve KS eigenvalue problem" step of the self-consistent procedure in Fig. 2.1 in the case the central region calculation using the Green's function method. Notice that in order to obtain $\mathbf{D}^{\mathrm{eq}}$ and $\mathbf{D}^{\mathrm{neq}}$ one has to compute $\mathbf{\Sigma}_L$,$\mathbf{\Sigma}_R$, and $\mathbf{G}_C^r$ for many different energy points E.

of the electrodes typically consists of relatively few atoms and obtaining the correspondingly small electrode Hamiltonians in this way is quite fast.

The evaluation of the central region Hamiltonian $\mathbf{H}_C$ follows a slightly modified flow chart, as depicted in Fig. 3.5, in order to incorporate the Green's function description presented in the previous sections. Compared to the molecule and bulk calculations, the "solve KS eigenvalue problem" step of the self-consistent procedure is different. Now the eigenproblem solution is obtained by means of the Green's function matrix $\mathbf{G}_C$ and the density matrix $\mathbf{D}$ (from which the density $n(\mathbf{r})$ is available, see Eq. (2.18)). We have already shown how perform the first two steps in this modified flow diagram. Let us now describe the step of obtaining the density matrix $\mathbf{D}$ from $\mathbf{G}_C$ in two important cases.

Equilibrium case $\mu_L = \mu_R \equiv \mu$

We can use the following integral expression from NEGF theory (derived in Sec. B.5) to calculate the electron density matrix

$$\mathbf{D} = -\frac{1}{\pi}\int_{-\infty}^{\infty} \mathrm{Im}\{\mathbf{G}_C\} f(E-\mu)\,\mathrm{d}E, \tag{3.46}$$

under equilibrium conditions. In practice the integral is bounded from below by the bottom valence-band edge and from above by the vanishing of the Fermi function when $E > \mu$. The only problem with Eq. (3.46) is therefore that the Green's function is in general a rapidly varying function along the real axis. This implies that an accurate determination of the integral requires many energy points, often more than 5000. Fortunately, one can get around this issue by using complex functions theory. Since the Green's function is an analytical function and can be extended into the complex plane, we can evaluate the integral in Eq. (3.46) according to the residue theorem by integrating along a

contour in the complex plane, which encloses the poles of the Fermi function. A detailed description of this contour technique can be found in Ref. [39]. In the complex plane, the functions are smooth and only a few points are needed for very accurate integration, usually around 50 points. This factor 100 reduction in the computational expense is one of the most important virtues of the Green's function method as a practical approach [37].

Non-equilibrium case with finite bias $|\mu_L - \mu_R| > 0$

In the non-equilibrium case we have to use the general NEGF expression

$$\mathbf{D} = \frac{1}{2\pi\hat{\imath}} \int_{-\infty}^{\infty} \mathbf{G}^{<} \, \mathrm{d}E. \tag{3.47}$$

to obtain the density matrix. The "lesser" Green's function matrix is given by

$$\mathbf{G}^{<} = \hat{\imath}\mathbf{G}_C\Big(\mathbf{\Gamma}_L f(E-\mu_L) + \mathbf{\Gamma}_R f(E-\mu_R)\Big)\mathbf{G}_C^{\dagger}, \tag{3.48}$$

where $\mathbf{\Gamma}_L$ and $\mathbf{\Gamma}_R$ are the broadening matrices defined in Eq. (3.43) (Eqs. (3.47) and (3.48) are also derived in Sec. B.5). We note that, in practice, only the first and last block columns of $\mathbf{G}_C$ are actually needed in Eq. (3.48) because of the zero parts of $\mathbf{\Gamma}_L$ and $\mathbf{\Gamma}_R$ (same arguments as used in Sec. 3.2.4). However, unlike the retarded Green's function $\mathbf{G}_C$, the lesser Green's function $\mathbf{G}^{<}$ is not analytic away from the real energy axis, which makes it impossible to apply the contour method. Fortunately, it is possible to rewrite Eq. (3.47) as two separate contributions, given by

$$\mathbf{D} = \mathbf{D}^{\mathrm{eq}} + \mathbf{D}^{\mathrm{neq}}, \tag{3.49}$$

where the first contribution is

$$\begin{aligned} \mathbf{D}^{\mathrm{eq}} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Big(\mathbf{G}_C\mathbf{\Gamma}_L\mathbf{G}_C^{\dagger} + \mathbf{G}_C\mathbf{\Gamma}_R\mathbf{G}_C^{\dagger}\Big) f(E-\mu) \, \mathrm{d}E \\ &= -\frac{1}{\pi} \int_{-\infty}^{\infty} \mathrm{Im}\{\mathbf{G}_C\} f(E-\mu) \, \mathrm{d}E, \end{aligned} \tag{3.50}$$

which is identical to the equilibrium case in Eq. (3.46), and hence obtained using the complex contour, and the second contribution is

$$\begin{aligned} \mathbf{D}^{\mathrm{neq}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Big(&\mathbf{G}_C\mathbf{\Gamma}_L\mathbf{G}_C^{\dagger}\big(f(E-\mu_L) - f(E-\mu)\big) \\ &+ \mathbf{G}_C\mathbf{\Gamma}_R\mathbf{G}_C^{\dagger}\big(f(E-\mu_R) - f(E-\mu)\big)\Big) \, \mathrm{d}E \end{aligned} \tag{3.51}$$

where the integrand is non-zero only over a limited range, for which $f(E-\mu)$ differs significantly from either $f(E-\mu_L)$ or $f(E-\mu_R)$. This range is of the order of the bias window $\mu_L \leq E \leq \mu_R$, for $\mu_L < \mu_R$, but has to be evaluated along

the real energy axis with relatively small steps. For large biases the calculation of the non-equilibrium density matrix $\mathbf{D}^{\text{neq}}$ is the overall most time consuming part of the Green's function method (see benchmark results in Sec. 3.5.1).

We note in passing that if we choose $\mu = \mu_L$ in Eq. (3.51), then the first term of the integrand cancels out. Similarly, choosing $\mu = \mu_R$ eliminates the second term. This means that we have two simple and equivalent formulas by which to obtain the density matrix $\mathbf{D}$ in the non-equilibrium case. In practice, however, the results from the $\mu = \mu_L$ calculation and the $\mu = \mu_R$ calculation often differs, in particular because of errors when evaluating $\mathbf{D}^{\text{neq}}$. This difference can then be used to estimate the numerical integration error, and also to combine the two results as a weighted sum to improve the final accuracy (see Ref. [39]).

3.3 Wave function matching method

This section is devoted to the solution of the two-probe Schrödinger equation in Eq. (3.9) using the wave function matching (WFM) method. This method has become increasingly popular in recent years mainly because it extends naturally and in a transparent fashion the perception of electrons being transmitted and reflected, much in the spirit of the original Landauer picture (Sec. 3.1.2). As such, the basic approach of the WFM method has been used under different names by different people; the scattering states approach [39, 40], the mode matching method [59, 55], the over-bridging boundary-matching scheme [33, 60], and the WFM method [52, 61, 62]. Here we will adopt the last name and the formalism corresponding to it. It was initially developed by Ando [52] and subsequently refined by Brocks *et al.* [Ref. [62] and references therein]. In Chap. 5 we will develop the WFM method even further to achieve a form that is much more efficient in practice (up to an order of magnitude faster). The formal equivalence between the WFM method and the Green's function method that was presented in the previous section has been proven in [55].

3.3.1 Scattering wave function for two-probe systems

As we saw in the previous sections, the conductance and current through a nano-scale device attached to two reservoirs is proportional to the quantum-mechanical probability $T(E)$ that an incoming electron at energy E in the one reservoir will transmit to the other reservoir. In the following we set out to find $T(E)$ in terms of the individual transmission probabilities t_{ij} that enters in the Landauer formula Eq. (3.6), and also the density matrix $\mathbf{D}$ of the central region, by solving the infinite Schrödinger equation $(E\mathbf{S} - \mathbf{H})\mathbf{c} = \mathbf{0}$ defined in Eq. (3.9) directly for the scattered wave function $\mathbf{c}$.

The starting point is the same two-probe systems as before (see Sec. 3.1.4). In particular, we assume that the infinite structure is divided into principal layers numbered $i = -\infty, \ldots, \infty$ and composed of a finite central (C) region

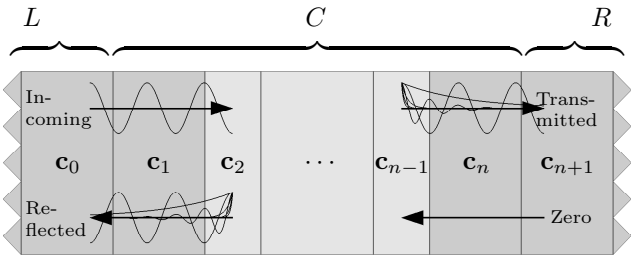


Figure 3.6: Schematic representation of WFM applied to layered two-probe systems, where the central device region, consisting of layers $i = 1, \ldots, n$, is attached to left and right semi-infinite electrodes. The incoming propagating state from the left electrode is scattered in the central region and end up as reflected and transmitted superpositions of propagating and evanescent states.

containing the device and two semi-infinite left (L) and right (R) electrode regions. The Schrödinger equation to be solved can then be written (see Fig. 3.6)

$$\begin{pmatrix} \ddots & \ddots & & & & & \\ \ddots & \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} & & & & \\ & \bar{\mathbf{H}}_{L,L}^\dagger & \begin{pmatrix} & & \\ & \bar{\mathbf{H}}_C & \\ & & \end{pmatrix} & & \bar{\mathbf{H}}_{R,R} & & \\ & & & \bar{\mathbf{H}}_{R,R}^\dagger & \bar{\mathbf{H}}_R & \ddots & \\ & & & & \ddots & \ddots & \end{pmatrix} \begin{pmatrix} \vdots \\ \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_n \\ \mathbf{c}_{n+1} \\ \vdots \end{pmatrix} = \mathbf{0}, \quad (3.52)$$

where the finite matrix of the central device $\bar{\mathbf{H}}_C$ is given in Eq. (3.12) using the usual notation $\bar{\mathbf{H}} \equiv E\mathbf{S} - \mathbf{H}$, and the wave function in layer i is represented by a column vector $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,m_i})^T$ of the expansion coefficients, where m_i is the number of orbitals in the layer. Thus the wave function $\mathbf{c}$ extending over the entire system is written as $\mathbf{c} = (\mathbf{c}_{-\infty}^{\mathrm{T}}, \ldots, \mathbf{c}_{\infty}^{\mathrm{T}})^{\mathrm{T}}$. Notice that we also in this setup assume that the border layers 1 and n of the central region are identical to a layer of the connecting electrodes. We will get back to this later.

As is known from scattering theory, the solution to Eq. (3.52) for the part far deep in the left electrode will correspond to a wave which is a superposition of the Bloch modes of this bulk material. A Bloch mode in this context is a wave that propagates in an ideal periodic lattice without loss (we will determine these in the next section). Assuming then, that an electron is incident from the left electrode, it will always be propagating in one of the right-going Bloch modes, say number i, available there. This is shown as the incoming wave in Fig. 3.6.

Subsequently, when it hits the central device region, some of the amplitude of the electron wave will be reflected back into the various left-going modes of the left electrode. These can be both Bloch modes and so-called evanescent modes, which decay exponentially. Alternatively, the electron may tunnel all the way through the device region, in which case some of the electron wave is transmitted into the right-going modes, denoted for example by j, of the right electrode. By finding the scattering wave function $\mathbf{c}$ corresponding to this situation it is then possible to determine how much was transmitted, yielding t_{ij}. This can be done by directly matching the wave functions in layers $\mathbf{c}_{-1}$ and $\mathbf{c}_{n+1}$ to the corresponding first and last layer of the central region. Subsequently $T(E)$ can be found by summing all contributions t_{ij} via the Landauer-Büttiker expression Eq. (3.8). We will now present the computational aspects of this technique, which is the core of the WFM method.

3.3.2 Bulk modes of the electrodes

Initially we need to calculate the modes of the bulk electrodes. To this end we look at the Schrödinger equation for the ideal left electrode, given by

$$\begin{pmatrix} \ddots & \ddots & & & \\ \ddots & \bar{\boldsymbol{H}}_L & \bar{\boldsymbol{H}}_{L,L} & & \\ & \bar{\boldsymbol{H}}_{L,L}^\dagger & \bar{\boldsymbol{H}}_L & \bar{\boldsymbol{H}}_{L,L} & \\ & & \bar{\boldsymbol{H}}_{L,L}^\dagger & \bar{\boldsymbol{H}}_L & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ \mathbf{c}_{L,i-1} \\ \mathbf{c}_{L,i} \\ \mathbf{c}_{L,i+1} \\ \vdots \end{pmatrix} = \mathbf{0}, \tag{3.53}$$

which has a block-Toeplitz structure and therefore can be written simply as

$$\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{c}_{L,i-1} + \bar{\mathbf{H}}_L \mathbf{c}_{L,i} + \bar{\mathbf{H}}_{L,L} \mathbf{c}_{L,i+1} = \mathbf{0}, \tag{3.54}$$

for $i = -\infty, \ldots, \infty$. The subscript L on the solution $\mathbf{c}_L$ designates that this is for the left electrode. Then, since this system is infinitely periodic, Bloch's theorem predicts that the $\mathbf{c}_{L,i}$'s differ only by a phase factor [2], i.e.,

$$\mathbf{c}_{L,i} = e^{i\mathrm{kd}} \mathbf{c}_{L,i-1}, \tag{3.55}$$

where k is the wave number and d is the distance between layers. Therefore, by defining $\lambda_L \equiv e^{i\mathrm{kd}}$ and using Eqs. (3.54) and (3.55), we can isolate a simple expression for $\mathbf{c}_{L,i}$, given by

$$\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{c}_{L,i} + \lambda_L \bar{\mathbf{H}}_L \mathbf{c}_{L,i} + \lambda_L^2 \bar{\mathbf{H}}_{L,L} \mathbf{c}_{L,i} = \mathbf{0}, \tag{3.56}$$

which is a quadratic eigenvalue problem (QEP).

In principle, the simplest way to solve Eq. (3.56) is by finding the eigenvalues and eigenvectors of the so-called transfer matrix[4]

$$\mathbf{T}_L = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\bar{\mathbf{H}}_{L,L}^{-1}\bar{\mathbf{H}}_{L,L}^{\dagger} & \bar{\mathbf{H}}_{L,L}^{-1}\bar{\mathbf{H}}_{L} \end{pmatrix} \tag{3.57}$$

where $\mathbf{I}$ is the identity $m_L \times m_L$ matrix. However, such a solution obviously requires that $\bar{\mathbf{H}}_{L,L}$ can be inverted which is seldom the case. In systems where $\bar{\mathbf{H}}_{L,L}$ is singular or ill-conditioned (all systems in this work) it is then most convenient to linearize the QEP and solve the equivalent generalized eigenvalue problem [63]

$$\mathbf{A}\begin{pmatrix} \mathbf{c}_{L,i} \\ \mathbf{c}_{L,i+1} \end{pmatrix} = \lambda_L \mathbf{B}\begin{pmatrix} \mathbf{c}_{L,i} \\ \mathbf{c}_{L,i+1} \end{pmatrix} \tag{3.58}$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\bar{\mathbf{H}}_{L,L} & -\bar{\mathbf{H}}_{L} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{H}}_{L,L} \end{pmatrix}, \tag{3.59}$$

which is of order $2m_L \times 2m_L$ and gives the $2m_L$ solutions of Eq. (3.56).

We can implement Eq. (3.58) straightforwardly by calling the LAPACK routines DGGEV/ZGGEV [29] for the real/complex case. However, it is also possible to apply a simple "shift-and-invert" trick to take advantage of the zero and identity subblocks of $\mathbf{A}$ and $\mathbf{B}$. See the details of this approach in App. C.

Since the $\mathbf{c}_{L,i}$ vectors for different layers i are related simply via Eq. (3.55) we will from here on skip the implied layer subscript i. Let us instead designate the $2m_L$ solutions of Eq. (3.58) by numbers $k = 1, \ldots, 2m_L$. Solving Eq. (3.58) thus provides the electrode modes $\mathbf{c}_{L,k}$ (as the eigenvectors), and the phase factors $\lambda_{L,k}$ (as eigenvalues), which we from here on refer to as Bloch factors. In general, the eigenvalues appear as complex pairs $(\lambda_{L,k}^{+}, \lambda_{L,k}^{-})$, related by $\lambda_{L,k}^{+}\lambda_{L,k}^{-} = 1$ [63], which in our case indicates that half the modes are right-going (+) and half are left-going (−), as can be expected from symmetry arguments.

A simple algorithm running through the $2m_L$ solutions one by one can then be used to group the modes into three categories:

1. Trivial solutions which have eigenvalues $|\lambda_{L,k}^{+}| = 0$ or $|\lambda_{L,k}^{-}| = \texttt{inf}$, which are interpreted as unphysical or very rapidly decaying or growing modes. These eigenpairs are discarded.

2. The so-called Bloch modes that have real wave numbers k giving Bloch factors with $|\lambda_{L,k}| = 1$, and correspond to propagating waves that go from one layer to the next with constant amplitude.

[4]The transfer matrix used here is defined $\begin{pmatrix} \mathbf{c}_{L,i} \\ \mathbf{c}_{L,i+1} \end{pmatrix} = \mathbf{T}_L \begin{pmatrix} \mathbf{c}_{L,i-1} \\ \mathbf{c}_{L,i} \end{pmatrix}$.

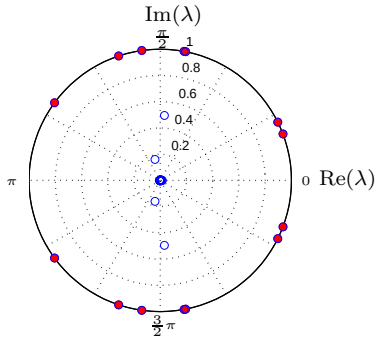


Figure 3.7: A polar plot showing the positions of the 243 complex eigenvalues (blue/circles) inside the unit disc (i.e., $|\lambda| \leq 1$) for an Au(111) electrode with 27 atoms per unit cell at $E = -2$ eV. There are 21 eigenvalues corresponding to Bloch modes (red/filled dots) which are located on the unit circle.

3. The remaining modes that have complex k and $|\lambda_{L,k}| \neq 1$ and represent evanescent (exponentially growing or decaying) waves.[5]

A polar plot of the Bloch factors with $|\lambda_k| \leq 1$ of an example Au(111) electrode is shown in Fig. 3.7. The few Bloch factors that correspond to the propagating Bloch modes are highlighted.

We note that, in practice, it makes sense to keep only those eigenpairs that have eigenvalues within the intervals [55]

$$\lambda_{\min} \leq |\lambda^{+}_{L,k}| \leq 1 \quad \text{and} \quad 1 \leq |\lambda^{-}_{L,k}| \leq \lambda^{-1}_{\min}, \tag{3.60}$$

for a reasonable choice of $\lambda_{\min}$. Modes with $|\lambda^{\pm}_{L,k}|$ outside these intervals are decaying or growing very rapidly and their influence in an actual calculation is in most cases insignificant. What exactly constitutes a reasonable choice of the parameter $\lambda_{\min}$ is a key subject of Chap. 5.

Finally, let us assume that the number of modes selected from category 2 and 3 is $2\mathring{m}_L$ out of the total $2m_L$ solutions. In the following it is then convenient to classify these modes according to the propagation direction. In the case of Bloch modes, we calculate the group velocity from the expression [59]

$$v_{L,k} = -\frac{2\mathrm{d}}{\hbar}\mathrm{Im}\left\{\lambda_{L,k}\mathbf{c}^{\dagger}_{L,k}\bar{\mathbf{H}}^{\dagger}_{L,L}\mathbf{c}_{L,k}\right\}, \tag{3.61}$$

and use the sign of $v_{L,k}$ to distinguish whether they are right-going (+) or left-going (−). For the evanescent modes, we know by definition (see Eq. (3.55))

[5] Even though the evanescent modes do not contribute to the transmission in the ideal unbounded electrode, they will provide important exponential tails leaving and entering "impurities" such as a scatter region or boundaries.

that we have $|\lambda^{+}_{L,k}| < 1$ for right-going modes and $|\lambda^{-}_{L,k}| > 1$ for left-going modes.

The mode selection and classification for the left electrode leads to the definition of two mode matrices:

$$\mathbf{C}^{\pm}_{L} = \left(\mathbf{c}^{\pm}_{L,1}, \ldots, \mathbf{c}^{\pm}_{L,\mathring{m}_L}\right), \tag{3.62}$$

which has the right-going (+) and left-going (−) coefficient vectors as columns, and two corresponding diagonal matrices

$$\mathbf{\Lambda}^{\pm}_{L} = \mathrm{diag}\{\lambda^{\pm}_{1}, \lambda^{\pm}_{2}, \ldots, \lambda^{\pm}_{\mathring{m}_L}\}, \tag{3.63}$$

that contain the Bloch factors. We also note, without explicit derivation, that similar considerations for the right electrode would lead to definitions of matrices $\mathbf{C}^{\pm}_{R}$ and $\mathbf{\Lambda}^{\pm}_{R}$, built from solutions of the corresponding QEP for the right system.

3.3.3 Block tridiagonal system of linear equations

Returning now to the infinite two-probe Schrödinger equation in Eq. (3.9), we know that the solution to such a second-order differential equation can be determined be specifying a value and a first derivative at one point or by specifying values at two points. In our case, we will do the latter as follows.

As discussed in Sec. 3.3.1 we can describe the incident, reflected and transmitted parts of scattering wave function as superpositions of bulk modes. Thus it is possible to express the wave functions in layers $i = 0$ and $i = \mathrm{n}+1$ in terms of the mode matrices, for example as

$$\mathbf{c}_0 = \mathbf{C}^{+}_{L}\mathbf{a}^{\mathrm{in}} + \mathbf{C}^{-}_{L}\mathbf{a}^{\mathrm{ref}}, \quad \mathbf{c}_{\mathrm{n}+1} = \mathbf{C}^{+}_{R}\mathbf{a}^{\mathrm{trans}}, \tag{3.64}$$

where $\mathbf{a}^{\mathrm{in}}$,$\mathbf{a}^{\mathrm{ref}}$, and $\mathbf{a}^{\mathrm{trans}}$ are vectors holding the expansion coefficients in the basis of bulk modes (which is be assumed to be complete [56]). We note that the left-going part of $\mathbf{c}_{\mathrm{n}+1}$ is fixed to $\mathbf{0}$ from the outset. Also, in order to specify a particular incident right-going mode k we can simply set $[\mathbf{a}^{\mathrm{in}}]_i = \delta_{i,k}$.

Now we may take the *L–C–R* splitting of our two-probe systems into account, i.e., that the boundaries were chosen far enough into the semi-infinite electrodes layers, that the wave functions for layers $i < 1$ and $i > \mathrm{n}$ correspond to the ideal electrode case, for which $\mathbf{c}_i = \lambda\mathbf{c}_{i-1}$ is valid. In our notation this allows us to write the wave functions in layers $i = -1$ and $i = \mathrm{n}+2$, as

$$\mathbf{c}_{-1} = \mathbf{C}^{+}_{L}(\mathbf{\Lambda}^{+}_{L})^{-1}\mathbf{a}^{\mathrm{in}} + \mathbf{C}^{-}_{L}(\mathbf{\Lambda}^{-}_{L})^{-1}\mathbf{a}^{\mathrm{ref}}, \quad \mathbf{c}_{\mathrm{n}+2} = \mathbf{C}^{+}_{R}\mathbf{\Lambda}^{+}_{R}\mathbf{a}^{\mathrm{trans}}. \tag{3.65}$$

by using both the mode matrices and the Bloch factor matrices. Inserting $\mathbf{c}_{-1}$

and $\mathbf{c}_{\mathrm{n}+2}$ in Eq. (3.52), i.e., specifying two values of the solution, we arrive at

$$\begin{pmatrix} \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} & & \\ \bar{\mathbf{H}}_{L,L}^\dagger & \begin{pmatrix} & \\ \bar{\mathbf{H}}_C & \\ & \end{pmatrix} & & \bar{\mathbf{H}}_{R,R} \\ & & \bar{\mathbf{H}}_{R,R}^\dagger & \bar{\mathbf{H}}_R \end{pmatrix} \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_n \\ \mathbf{c}_{n+1} \end{pmatrix} = \begin{pmatrix} -\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{c}_{-1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\bar{\mathbf{H}}_{L,L}\mathbf{c}_{\mathrm{n}+2} \end{pmatrix}, \tag{3.66}$$

which is the finite system of linear equations to be solved in the WFM method.

In order to implement an appropriate WFM algorithm we first rewrite the linear system in Eq. (3.66) in a more convenient form. Considering Eqs. (3.64) and (3.65) we see that it is possible to express $\mathbf{c}_{-1}$ in terms of $\mathbf{c}_0$ and the known coefficients $\mathbf{a}^{\mathrm{in}}$ of the incoming wave, if matrices $\mathbf{C}_L^{\pm}$ and $(\mathbf{\Lambda}_L^{\pm})^{-1}$ were to commute.[6] Let us assume that we can invert the mode matrices and write $\tilde{\mathbf{C}}_L^{\pm}\mathbf{C}_L^{\pm} = \mathbf{I}$, where $\tilde{\mathbf{C}}_L^{\pm}$ are the so-called (Moore-Penrose) pseudo-inverses [1]. Consequently, inserting this in Eq. (3.65) and using Eq. (3.64) yields

$$\begin{aligned} \mathbf{c}_{-1} &= \mathbf{C}_L^{+}(\mathbf{\Lambda}_L^{+})^{-1}\tilde{\mathbf{C}}_L^{+}\mathbf{C}_L^{+}\mathbf{a}^{\mathrm{in}} + \mathbf{C}_L^{-}(\mathbf{\Lambda}_L^{-})^{-1}\tilde{\mathbf{C}}_L^{-}\mathbf{C}_L^{-}\mathbf{a}^{\mathrm{ref}} \\ &= \left[(\mathbf{B}_L^{+})^{-1} - (\mathbf{B}_L^{-})^{-1}\right]\mathbf{C}_L^{+}\mathbf{a}^{\mathrm{in}} + (\mathbf{B}_L^{-})^{-1}\mathbf{c}_0, \end{aligned} \tag{3.67}$$

where the so-called Bloch matrices [55] for the left electrode

$$\mathbf{B}_L^{\pm} = \mathbf{C}_L^{\pm}\mathbf{\Lambda}_L^{\pm}\tilde{\mathbf{C}}_L^{\pm} \tag{3.68}$$

have been introduced. In the same manner, we can express $\mathbf{c}_{\mathrm{n}+2}$ in terms of $\mathbf{c}_{\mathrm{n}+1}$ by inserting $\mathbf{I} = \tilde{\mathbf{C}}_R^{+}\mathbf{C}_R^{+}$ appropriately, i.e.

$$\mathbf{c}_{\mathrm{n}+2} = \mathbf{C}_R^{+}\mathbf{\Lambda}_R^{+}\tilde{\mathbf{C}}_R^{+}\mathbf{C}_R^{+}\mathbf{a}^{\mathrm{trans}} = \mathbf{B}_R^{+}\mathbf{c}_{\mathrm{n}+1}, \tag{3.69}$$

where

$$\mathbf{B}_R^{\pm} = \mathbf{C}_R^{\pm}\mathbf{\Lambda}_R^{\pm}\tilde{\mathbf{C}}_R^{\pm} \tag{3.70}$$

are the Bloch matrices for the right electrode. Finally, the combination of Eqs. (3.66), (3.67), and (3.69) results in the block tridiagonal system of linear equations, given by

$$\begin{pmatrix} \mathbf{g}_L^{-1} & \bar{\mathbf{H}}_{L,L} & & & \\ \bar{\mathbf{H}}_{L,L}^\dagger & \bar{\mathbf{H}}_1 & \bar{\mathbf{H}}_{1,2} & & \\ & \bar{\mathbf{H}}_{1,2}^\dagger & \ddots & \ddots & \\ & & \ddots & \bar{\mathbf{H}}_{\mathrm{n}} & \bar{\mathbf{H}}_{R,R} \\ & & & \bar{\mathbf{H}}_{R,R}^\dagger & \mathbf{g}_R^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{\mathrm{n}} \\ \mathbf{c}_{\mathrm{n}+1} \end{pmatrix} = \begin{pmatrix} \mathbf{q}_0 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \tag{3.71}$$

[6] This would only be the case for orthonormal $\mathbf{c}_{L,k}$ vectors, i.e., $\mathbf{C}_L^{\pm\dagger}\mathbf{C}_L^{\pm} = \mathbf{I}$, which is not a valid assumption in our derivation.

where

$$\mathbf{g}_L = [\bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}^\dagger (\mathbf{B}_L^-)^{-1}]^{-1}, \quad \mathbf{g}_R = [\bar{\mathbf{H}}_R + \bar{\mathbf{H}}_{R,R}\mathbf{B}_R^+]^{-1} \tag{3.72}$$

and the remaining source term $\mathbf{q}_0$ on the right is

$$\mathbf{q}_0 = \bar{\mathbf{H}}_{L,L}^\dagger \left[(\mathbf{B}_L^-)^{-1} - (\mathbf{B}_L^+)^{-1}\right] \mathbf{C}_L^+ \mathbf{a}^{\text{in}}. \tag{3.73}$$

We now present how to implement the solution of Eq. (3.71) in a straightforward and efficient manner. Initially we need to determine the pseudo-inverses of $\mathbf{C}_L^\pm$ and $\mathbf{C}_R^\pm$ in order to obtain the Bloch matrices $\mathbf{B}_L^\pm$ and $\mathbf{B}_R^\pm$. The numerically most stable way to proceed is to perform a QR factorization of the mode matrices (assuming that these have full rank $\mathring{m}_L$), written as

$$\mathbf{C}_L^\pm = \mathbf{Q}_L^\pm \mathbf{R}_L^\pm, \quad (\mathbf{Q}_L^\pm)^\dagger \mathbf{Q}_L^\pm = \mathbf{I}, \tag{3.74}$$

for the left electrode and similarly with $L \to R$ for the right electrode. Here $\mathbf{R}_L^\pm$ are upper triangular $\mathring{m}_L \times \mathring{m}_L$ matrices and $\mathbf{Q}_L^\pm$ are orthogonal $m_L \times \mathring{m}_L$ matrices. It is then easy to form the pseudo-inverses by solving

$$\mathbf{R}_L^\pm \tilde{\mathbf{C}}_L^\pm = (\mathbf{Q}_L^\pm)^\dagger, \tag{3.75}$$

which requires only back substitution since the $\mathbf{R}_L^\pm$ matrices are already upper triangular. The same operations hold for $L \to R$ and the Bloch matrices can subsequently be formed directly from Eqs. (3.68) and (3.70).

We are now in a position to evaluate $\mathbf{g}_L^{-1}$, $\mathbf{g}_R^{-1}$, and $\mathbf{q}_0$, and solve the linear system of equations using the block Gaussian elimination sweeps. Performing the elimination of the lower off-diagonal blocks of Eq. (3.71) with a downwards sweep (see Eqs. (1.4)–(1.5)) yields

$$\begin{pmatrix} \bar{\mathbf{H}}_0' & \bar{\mathbf{H}}_{L,L} & & & \\ & \bar{\mathbf{H}}_1' & \bar{\mathbf{H}}_{1,2} & & \\ & & \ddots & \ddots & \\ & & & \bar{\mathbf{H}}_{\mathrm{n}}' & \bar{\mathbf{H}}_{R,R} \\ & & & & \bar{\mathbf{H}}_{\mathrm{n}+1}' \end{pmatrix} \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{\mathrm{n}} \\ \mathbf{c}_{\mathrm{n}+1} \end{pmatrix} = \begin{pmatrix} \mathbf{q}_0' \\ \mathbf{q}_1' \\ \vdots \\ \mathbf{q}_{\mathbf{n}}' \\ \mathbf{q}_{\mathbf{n}+1}' \end{pmatrix}, \tag{3.76}$$

where $\bar{\mathbf{H}}_0' = \mathbf{g}_L^{-1}$, $\mathbf{q}_0' = \mathbf{q}_0$, and

$$\bar{\mathbf{H}}_i' = \bar{\mathbf{H}}_i - \bar{\mathbf{H}}_{i,i-1}(\bar{\mathbf{H}}_{i-1}')^{-1}\bar{\mathbf{H}}_{i-1,i}, \quad i = 1, \ldots, \mathrm{n}+1, \tag{3.77}$$

$$\mathbf{q}_i' = -\bar{\mathbf{H}}_{i,i-1}(\bar{\mathbf{H}}_{i-1}')^{-1}\mathbf{q}_{i-1}', \quad i = 1, \ldots, \mathrm{n}+1, \tag{3.78}$$

using $\mathbf{H}_{\mathrm{n}+1} = \mathbf{g}_R^{-1}$ as initialization.

Succeeding these operations by an upwards sweep, then gives

$$\begin{pmatrix} \bar{\mathbf{H}}_0' & & & & \\ & \bar{\mathbf{H}}_1' & & & \\ & & \ddots & & \\ & & & \bar{\mathbf{H}}_{\mathrm{n}}' & \\ & & & & \bar{\mathbf{H}}_{\mathrm{n}+1}' \end{pmatrix} \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{\mathrm{n}} \\ \mathbf{c}_{\mathrm{n}+1} \end{pmatrix} = \begin{pmatrix} \mathbf{q}_0^\vee \\ \mathbf{q}_1^\vee \\ \vdots \\ \mathbf{q}_{\mathbf{n}}^\vee \\ \mathbf{q}_{\mathbf{n}+1}^\vee \end{pmatrix}, \tag{3.79}$$

where

$$\mathbf{q}_i^{\vee} = \mathbf{q}_i' - \bar{\mathbf{H}}_{i,i+1}(\bar{\mathbf{H}}_{i+1}')^{-1}\mathbf{q}_{i+1}^{\vee}, \quad i = 0, \ldots, \mathrm{n}. \tag{3.80}$$

with $\mathbf{q}_{\mathrm{n}+1}^{\vee} = \mathbf{q}_{\mathrm{n}+1}'$. Thus we end up with a block diagonal left-hand side matrix and can solve for the solutions $\mathbf{c}_i$ simply as

$$\mathbf{c}_i = (\bar{\mathbf{H}}_i')^{-1}\mathbf{q}_i^{\vee}, \quad i = 0, \ldots, \mathrm{n}+1, \tag{3.81}$$

which gives us the correctly matched scattering wave functions we were looking for, both inside the central region and in the connecting layers of the electrodes. We show in the next two sections how to benefit from the solutions $\mathbf{c}_i$.

In line with the above derivations we implement the WFM method in this thesis with the following $O(N)$ algorithm:

ALGORITHM V: *The WFM method*

1. obtain $\mathbf{C}_L^{\pm}, \mathbf{\Lambda}_L^{\pm}$ from QEP (Eqs. (3.56) − (3.63))
2. $[\mathbf{Q}_L^{\pm}, \mathbf{R}_L^{\pm}] = \mathrm{QR}\{\mathbf{C}_L^{\pm}\}$, solve $\mathbf{R}_L^{\pm}\tilde{\mathbf{C}}_L^{\pm} = (\mathbf{Q}_L^{\pm})^{\dagger}$
3. $\mathbf{B}_L^{\pm} := \mathbf{C}_L^{\pm}\mathbf{\Lambda}_L^{\pm}\tilde{\mathbf{C}}_L^{\pm}$
4. $\mathbf{g}_L^{-1} := \bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}^{\dagger}(\mathbf{B}_L^{-})^{-1}$

} the left elecrode

5. obtain $\mathbf{C}_R^{+}, \mathbf{\Lambda}_R^{+}$ from QEP (Eqs. (3.56) − (3.63))
6. $[\mathbf{Q}_R^{+}, \mathbf{R}_R^{+}] = \mathrm{QR}\{\mathbf{C}_R^{+}\}$, solve $\mathbf{R}_R^{+}\tilde{\mathbf{C}}_R^{+} = (\mathbf{Q}_R^{+})^{\dagger}$
7. $\mathbf{B}_R^{+} := \mathbf{C}_R^{+}\mathbf{\Lambda}_R^{+}\tilde{\mathbf{C}}_R^{+}$
8. $\mathbf{g}_R^{-1} := \bar{\mathbf{H}}_R + \bar{\mathbf{H}}_{R,R}\mathbf{B}_R^{+}$

} the right elecrode

9. $\mathbf{q}_0 := \bar{\mathbf{H}}_{L,L}^{\dagger}\left[(\mathbf{B}_L^{-})^{-1} - (\mathbf{B}_L^{+})^{-1}\right]\mathbf{C}_L^{+}\mathbf{a}^{\mathrm{in}}$
10. initialize $\bar{\mathbf{H}}_0' := \mathbf{g}_L^{-1}$, $\bar{\mathbf{H}}_{\mathrm{n}+1} := \mathbf{g}_R^{-1}$, $\mathbf{q}_0' := \mathbf{q}_0$
11. **for** $i := 1, \ldots, \mathrm{n}+1$
12. solve $\bar{\mathbf{H}}_{i,i-1} = \mathbf{X}\bar{\mathbf{H}}_{i-1}'$ for $\mathbf{X}$
13. $\bar{\mathbf{H}}_i' := \bar{\mathbf{H}}_i - \mathbf{X}\bar{\mathbf{H}}_{i-1,i}$
14. $\mathbf{q}_i' := -\mathbf{X}\mathbf{q}_{i-1}'$
15. **end**

} downwards sweep

16. initialize $\mathbf{q}_{\mathrm{n}+1}^{\vee} := \mathbf{q}_{\mathrm{n}+1}'$
17. **for** $i := \mathrm{n}, \ldots, 0$
18. solve $\bar{\mathbf{H}}_{i,i+1} = \mathbf{X}\bar{\mathbf{H}}_{i+1}'$ for $\mathbf{X}$
19. $\mathbf{q}_i^{\vee} := \mathbf{q}_i' - \mathbf{X}\mathbf{q}_{i+1}^{\vee}$
20. **end**

} upwards sweep

21. **for** $i := 0, \ldots, \mathrm{n}+1$
22. solve $\bar{\mathbf{H}}_i'\mathbf{c}_i = \mathbf{q}_i^{\vee}$
23. **end**

} scattering states solution

(3.82)

Notice that we can calculate for all the possible incident modes simultaneously by using $\mathbf{a}^{\text{in}} \to \mathbf{I}$, where $\mathbf{I}$ is of order $\mathring{m}_L \times \mathring{m}_L$. Furthermore, we have in this algorithm used that $\mathbf{B}^{-1} = (\mathbf{C}\mathbf{\Lambda}\tilde{\mathbf{C}})^{-1} = \mathbf{C}\mathbf{\Lambda}^{-1}\tilde{\mathbf{C}} = \mathbf{C}\ \mathrm{diag}\{\lambda_1^{-1}, \ldots, \lambda_m^{-1}\}\tilde{\mathbf{C}}$, with all super and subscripts implied, in order to avoid any explicit inversion of the Bloch matrices $\mathbf{B}_L^{\pm}$ and $\mathbf{B}_R^{\pm}$, which would be inefficient.

3.3.4 Transmission calculations

As a final important aspect of the WFM approach we want to determine the transmission coefficients t_{ij} in order to obtain the total transmission $T(E)$ from the Landauer-Büttiker formula in Eq. (3.8). The transparency of the WFM description makes this quite easy. As we already know the specific mode that is incident, we simply have to relate the calculated scattering solution in the first layer of the right electrode $\mathbf{c}_{\mathrm{n}+1}$ to the right-going bulk modes available here.

More specifically, assume that the particular Bloch mode which is incident is available as column k of the mode matrix $\mathbf{C}_L^+$, i.e., using $[\mathbf{a}^{\text{in}}]_i = \delta_{i,k}$. We denote the resulting scattering wave function in layer $\mathrm{n}+1$ by $\mathbf{c}_{\mathrm{n}+1,k}$, which is calculated from the WFM linear system in Eq. (3.71). Then, we can write

$$\mathbf{C}_R^+ \mathbf{t}_k = \mathbf{c}_{\mathrm{n}+1,k}, \tag{3.83}$$

giving the kth column of $\mathbf{t}$, since $\mathbf{C}_R^+$ is the $m_R \times \mathring{m}_R$ column matrix holding the right-going bulk modes of the right electrode (and here assumed to have full rank). Thus to obtain the entire transmission matrix in one step, we simply solve

$$\mathbf{C}_R^+ \mathbf{t} = (\mathbf{c}_{\mathrm{n}+1,1}, \mathbf{c}_{\mathrm{n}+1,2}, \ldots, \mathbf{c}_{\mathrm{n}+1,\mathring{m}_L}), \tag{3.84}$$

where the right-hand side is obtained simultaneously using $\mathbf{a}^{\text{in}} \to \mathbf{I}$.

There are two additional points to consider before calculating $T(E)$ with the WFM method. First of all, there is no need to perform an upwards Gaussian elimination sweep as done in ALGORITHM V in the previous section. If we are only interested in the transmission coefficients t_{ij} (and not the reflection coefficients r_{ij}) then only $\mathbf{c}_{\mathrm{n}+1}$ is required, and this solution is readily available from the downwards sweep and one subsequent solve. In order to obtain the reflection coefficients, however, the $\mathbf{c}_0$ solution is needed, i.e.,

$$\mathbf{C}_L^- \mathbf{r} = (\mathbf{c}_{0,1}, \mathbf{c}_{0,2}, \ldots, \mathbf{c}_{0,\mathring{m}_L}) - \mathbf{C}_L^+, \tag{3.85}$$

which demands the full solution of the linear system. In this case the second term removes the components corresponding to the incident modes.

The other point to consider is that when using the Landauer-Büttiker formula in Eq. (3.8) it is assumed that the electrode modes carry unit current in the conduction direction. In other words, we have to weigh the transmission coefficients by the group velocities of the Bloch modes involved or "flux-normalize"

all the Bloch modes right from the beginning [64, 42]. In the current implementation we do the latter and require all the Bloch modes in $\mathbf{C}_L^{\pm}$ to satisfy

$$(\mathbf{c}_{L,k}^{\pm})^{\dagger}\mathbf{c}_{L,k}^{\pm} = \frac{\mathrm{d}_L}{v_{L,k}^{\pm}}, \quad \text{(Bloch modes)} \tag{3.86}$$

where $v_{L,k}^{\pm}$ are the group velocities as defined in Eq. (3.61) and d_L is the layer thickness of the left electrode. At the same time, all the evanescent electrode modes are (state-) normalized in the standard way, i.e.,

$$(\mathbf{c}_{L,k}^{\pm})^{\dagger}\mathbf{c}_{L,k}^{\pm} = 1, \quad \text{(Evanescent modes)} \tag{3.87}$$

and similarly in the case of the right electrode $L \to R$. The above normalizations ensure that $\mathbf{t}^{\dagger}\mathbf{t} + \mathbf{r}^{\dagger}\mathbf{r} = \mathbf{1}$ from the outset, which is also convenient for testing the accuracy of the obtained results, as we will see in Chap. 5.

Finally, the algorithm to calculate the total transmission in the WFM method can be written as:

ALGORITHM VI: *Calculate $T(E)$ using WFM*

1. initialize $\bar{\mathbf{H}}'_1 := \mathbf{g}_L^{-1},\ \bar{\mathbf{H}}_{\mathrm{n}+1} := \mathbf{g}_R^{-1},\ \mathbf{q}'_0 := \mathbf{q}_0$
2. **for** $i := 2, \ldots, \mathrm{n}+1$
3. solve $\bar{\mathbf{H}}_{i,i-1} = \mathbf{X}\bar{\mathbf{H}}'_{i-1}$ for $\mathbf{X}$
4. $\bar{\mathbf{H}}'_i := \bar{\mathbf{H}}_i - \mathbf{X}\bar{\mathbf{H}}_{i-1,i}$
5. $\mathbf{q}'_i := -\mathbf{X}\mathbf{q}'_{i-1}$
6. **end**
7. solve $\bar{\mathbf{H}}'_{\mathrm{n}+1}\mathbf{c}_{\mathrm{n}+1} = \mathbf{q}'_{\mathrm{n}+1}$
8. solve $\mathbf{C}_R^{+}\mathbf{t} = \mathbf{c}_{\mathrm{n}+1}$
9. obtain $T(E)$ from Eq. (3.8)

(3.88)

Here it is assumed that the matrices $\mathbf{C}_R^{+}$, $\mathbf{g}_L^{-1}$, $\mathbf{g}_R^{-1}$, and $\mathbf{q}_0$ (for $\mathbf{a}^{\mathrm{in}} \to \mathbf{I}$) have been provided by executing lines $1-9$ of ALGORITHM V. We note that this implementation uses the same number of solves and multiplications inside the sweep **for** loop as the Green's function approach in ALGORITHM IV. However, since the number of columns of $\mathbf{q}'_i$ (i.e., number of Bloch modes in the left electrode) is in general much lower than m_i (i.e., the total number of modes in the left electrode), the matrix-matrix multiplication in line 5 in ALGORITHM VI is relatively inexpensive compared to line 5 of ALGORITHM IV. This actually results in a notable speed-up in the downwards sweep performed (see Sec. 3.5.2).

3.3.5 Obtaining the self-consistent Hamiltonian

The correspondence between the WFM approach and the Landauer-Büttiker formalism allows us to merge the WFM method with the self-consistent DFT

procedure in the same manner as described in Sec. 3.1.5. More precisely, by using the obtained scattering states solutions for the central region,

$$\mathbf{c}_{C,k} = (\mathbf{c}_{1,k}^{\mathrm{T}}, \ldots, \mathbf{c}_{\mathbf{n},k}^{\mathrm{T}})^{\mathrm{T}}, \tag{3.89}$$

where k numbers the different incident Bloch modes either in the left electrode as described above, or in the right electrode under reverse circumstances, we are able to determine the density matrix directly as an outer product (cf. Eq. (3.14))

$$\mathrm{D}_{ij}^{\pm} = \sum_{k=1} [\mathbf{c}_{C,k}^{\pm}]_i [\mathbf{c}_{C,k}^{\pm}]_j^*, \tag{3.90}$$

where we have indicated by superscripts that the solutions $\mathbf{c}_{C,k}$ in Eq. (3.89) correspond to either the right-going (+) and left-going (−) case. Notice that there is an implicit energy dependence in this expression for $\mathbf{D}^{\pm}$ which is the electron energy E introduced from the outset via the notation $\bar{\mathbf{H}} \equiv E\mathbf{S} - \mathbf{H}$. It is subsequently possible to evaluate the electronic density $n(\mathbf{r})$ needed in the self-consistent procedure by means of the integral over energies in Eq. (3.15).

In this work, however, we will not adopt the WFM method as implemented in ALGORITHM VI in order to obtain the self-consistent Hamiltonian. Only for the transmission calculation is the method applied, i.e., without the upwards sweep, as explained in Sec. 3.3.4. Instead, we will use the combined method, presented in the next section, which is directly applicable as part of the self-consistent DFT procedure in the same manner as described in Sec. 3.2.5. With the new efficient scheme we develop in Chaps. (5) and (6) this turns out to be a very competitive approach especially for non-equilibrium cases.

3.4 Combining the two methods

As is apparent from the previous sections in this chapter, there are close similarities between the Green's function method and the WFM method within the modeling of quantum transport in two-probe systems. In particular, looking at the appropriate algorithms in Eqs. (3.40) and (3.82) we see that both have two main parts: A preliminary calculation for the bulk electrodes, and the solution for the central region by one downwards and one upwards block Gaussian elimination sweep. The formal equivalence between the two methods has already been established by Khomyakov *et al.* [55]. However, we are not aware of an actual implementation that attempts to combine the methods in practice. In this section we develop such a hybrid and the proper framework for its description.

3.4.1 Self-energy matrices revisited

To begin with we will derive an expression for the self-energy matrices defined in the Green's function formalism in terms of the quantities available in the

WFM approach. Consider again the linear system in Eq. (3.71), which has a block tridiagonal left-hand side matrix of $\mathrm{n}+2 \times \mathrm{n}+2$ blocks. We remind the reader that in out setup, the C region has an additional electrode layer at each boundary in order for $\bar{\mathbf{H}}_1 = \bar{\mathbf{H}}_L$ and $\bar{\mathbf{H}}_\mathrm{n} = \bar{\mathbf{H}}_R$ to hold (this is not the case in the previously published WFM formulations of Refs. [62] and [55]). Now notice that if we perform a block Gaussian elimination once (i.e., one row) from the top and once from the bottom in Eq. (3.71) and leave out the same lines, we get

$$\begin{pmatrix} \bar{\mathbf{H}}_1 - \mathbf{\Sigma}_L & \bar{\mathbf{H}}_{1,2} & & \\ \bar{\mathbf{H}}_{1,2}^\dagger & \bar{\mathbf{H}}_2 & \ddots & \\ & \ddots & \ddots & \bar{\mathbf{H}}_{\mathrm{n}-1,\mathrm{n}} \\ & & \bar{\mathbf{H}}_{\mathrm{n}-1,\mathrm{n}}^\dagger & \mathbf{H}_\mathrm{n} - \mathbf{\Sigma}_R \end{pmatrix} \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_\mathrm{n} \end{pmatrix} = \begin{pmatrix} \mathbf{q}_1' \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \tag{3.91}$$

where

$$\mathbf{\Sigma}_L = \bar{\mathbf{H}}_{L,L}^\dagger [\bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}^\dagger (\mathbf{B}_L^-)^{-1}]^{-1} \bar{\mathbf{H}}_{L,L}, \tag{3.92}$$

and

$$\mathbf{\Sigma}_R = \bar{\mathbf{H}}_{R,R} [\bar{\mathbf{H}}_R + \bar{\mathbf{H}}_{R,R} \mathbf{B}_R^+]^{-1} \bar{\mathbf{H}}_{R,R}^\dagger, \tag{3.93}$$

are the self-energy matrices for the left and right electrode (using Eqs. (3.21), (3.22) and (3.72)), and $\mathbf{q}_1' = -\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{g}_L \mathbf{q}_0$. Moreover, $\mathbf{q}_1'$ can be easily rewritten as

$$\begin{aligned} \mathbf{q}_1' &= -\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{g}_L \bar{\mathbf{H}}_{L,L}^\dagger [(\mathbf{B}_L^-)^{-1} - (\mathbf{B}_L^+)^{-1}] \mathbf{C}_L^+ \mathbf{a}^\mathrm{in} \\ &= -\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{g}_L [\bar{\mathbf{H}}_{L,L}^\dagger (\mathbf{B}_L^-)^{-1} + \bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L} \mathbf{B}_L^+] \mathbf{C}_L^+ \mathbf{a}^\mathrm{in} \\ &= -[\bar{\mathbf{H}}_{L,L}^\dagger + \bar{\mathbf{H}}_{L,L}^\dagger \mathbf{g}_L \bar{\mathbf{H}}_{L,L} \mathbf{B}_L^+] \mathbf{C}_L^+ \mathbf{a}^\mathrm{in} \\ &= -[\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{C}_L^+ + \mathbf{\Sigma}_L \mathbf{C}_L^+ \mathbf{\Lambda}_L^+] \mathbf{a}^\mathrm{in} \end{aligned} \tag{3.94}$$

by taking Eqs. (3.73),(3.72), and (3.21) into account, and using that $\mathbf{B}_L^+$ satisfies $\bar{\mathbf{H}}_{L,L}^\dagger (\mathbf{B}_L^\pm)^{-1} + \bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L} \mathbf{B}_L^\pm = \mathbf{0}$ by definition (see Eq. (3.68)). Because of our particular setup, we can then obtain the solutions for the layers $\mathrm{n}+1$ and 0 as $\mathbf{c}_{\mathrm{n}+1,k} = \mathbf{B}_R^+ \mathbf{c}_{\mathrm{n},k}$ and $\mathbf{c}_{0,k} = (\mathbf{B}_L^-)^{-1}(\mathbf{c}_{1,k} - \lambda_{L,k}^+ \mathbf{c}_{L,k}^+) + \mathbf{c}_{L,k}^+$, and insert these in Eqs. (3.84) and (3.85), which simplifies to

$$\mathbf{C}_R^+ (\mathbf{\Lambda}_R^+)^{-1} \mathbf{t} = [\mathbf{c}_{\mathrm{n},1}, \mathbf{c}_{\mathrm{n},2}, \ldots, \mathbf{c}_{\mathrm{n},\hat{m}_L}], \tag{3.95}$$

and

$$\mathbf{C}_L^- \mathbf{\Lambda}_L^- \mathbf{r} = [\mathbf{c}_{1,1}, \mathbf{c}_{1,2}, \ldots, \mathbf{c}_{1,\hat{m}_L}] - \mathbf{C}_L^+ \mathbf{\Lambda}_L^+, \tag{3.96}$$

in order to obtain the scattering matrices $\mathbf{t}$ and $\mathbf{r}$ when using $\mathbf{a}^\mathrm{in} \rightarrow \mathbf{I}$.

More importantly, what has happened by doing the above block Gaussian elimination operations is that the Hamiltonian matrix entering the linear system in Eq. (3.91) is now identical to the inverse of the Green's function matrix $\mathbf{G}_C$ in Eq. (3.23). This implies that we have the simple relation

$$\mathbf{c}_i = [\mathbf{G}_C]_{i,1} \mathbf{q}_1', \quad i = 1, \ldots, \mathrm{n} \tag{3.97}$$

between the coefficients of the scattered wave in the central region and the first column blocks of the Green's function matrix. It also implies that we can calculate the self-energy matrices from Eqs. (3.92) and (3.93) and apply them in the Green's function approach. As seen from the benchmarks in Sec. 3.5 below, this is often faster than the usual calculation of the surface Green's functions $\mathbf{g}_L$ and $\mathbf{g}_R$ and the self-energy matrices using ALGORITHM II and Eqs. (3.21)–(3.22).

3.4.2 Hybrid method

Making a hybrid algorithm that calculates $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ using the WFM method and subsequently $\mathbf{G}_C$ in the central region (or $\mathbf{G}_{\mathbf{n},1}$ for transmission) using the Green's function method can be written:

ALGORITHM VII: *Hybrid method*

1. obtain $\mathbf{C}_L^-, \mathbf{\Lambda}_L^-$ from QEP (Eqs. (3.56) – (3.63))
2. $[\mathbf{Q}_L^-, \mathbf{R}_L^-] = \mathrm{QR}\{\mathbf{C}_L^-\}$, solve $\mathbf{R}_L^- \tilde{\mathbf{C}}_L^- = (\mathbf{Q}_L^-)^\dagger$
3. $\mathbf{B}_L^- := \mathbf{C}_L^- \mathbf{\Lambda}_L^- \tilde{\mathbf{C}}_L^-$
4. $\mathbf{\Sigma}_L := \bar{\mathbf{H}}_{L,L}^\dagger [\bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}^\dagger (\mathbf{B}_L^-)^{-1}]^{-1} \bar{\mathbf{H}}_{L,L}$

} the left electrode

5. obtain $\mathbf{C}_R^+, \mathbf{\Lambda}_R^+$ from QEP (Eqs. (3.56) – (3.63))
6. $[\mathbf{Q}_R^+, \mathbf{R}_R^+] = \mathrm{QR}\{\mathbf{C}_R^+\}$, solve $\mathbf{R}_R^+ \tilde{\mathbf{C}}_R^+ = (\mathbf{Q}_R^+)^\dagger$
7. $\mathbf{B}_R^+ := \mathbf{C}_R^+ \mathbf{\Lambda}_R^+ \tilde{\mathbf{C}}_R^+$
8. $\mathbf{\Sigma}_R := \bar{\mathbf{H}}_{R,R} [\bar{\mathbf{H}}_R + \bar{\mathbf{H}}_{R,R} \mathbf{B}_R^+]^{-1} \bar{\mathbf{H}}_{R,R}^\dagger$

} the right electrode

9. execute lines $5 - 26$ of ALGORITHM III or ALGORITHM IV.

(3.98)

Notice that here it suffices to determine only the left-going modes of the left electrode because only the self-energy matrices are obtained using the WFM method and not the source term $\mathbf{q}_1'$.

We would also like to point out that with this hybrid approach it is possible to have $\eta = 0$ (i.e., use real E) both for the solution of the QEP (Eqs. (3.56)–(3.63)) and the evaluation of the $\mathbf{G}_C$ in Eq. (3.23). As mentioned in Sec. 3.2.1, the η enforces the correct boundary conditions which, in the retarded case, is that right-going and left-going evanescent/Bloch waves decay/propagate towards ∞ and $-\infty$, respectively. This means that for nonzero η, the eigenvalues of the bulk QEPs with $|\lambda_k| = 1$ will be moved infinitesimally away from the unit circle to readily classify as either right or left-going modes. However, since we apply the group velocity formula in Eq. (3.61) to determine the propagation direction of the $|\lambda_k| = 1$ solutions, the infinitesimal η is not required for evaluating nor classifying the modes. In addition, it can be shown [4] that $\mathbf{G}_C|_{\eta \to 0^+} = \mathbf{G}_C|_{\eta \to 0^-}$ which makes $\eta = 0$ valid throughout ALGORITHM VII .

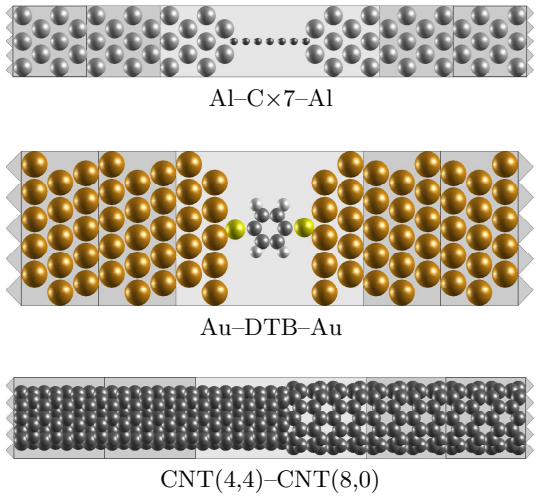


Figure 3.8: Example two-probe systems used for benchmarking.

3.5 Examples and benchmarks

To end this chapter we will present benchmarking results for the computational methods described in the above sections. Again we will use the ATK program as the existing baseline implementation, while the particular algorithms outlined (ALGORITHMS I-VII) have been implemented as additional functionality by this author. We will benchmark the methods on the example two-probe systems illustrated in Fig. 3.8. These systems are relatively small and can be readily investigated for their electronic transport properties on a single CPU. Furthermore, from looking at the literature we also note that these and similar systems are representative as *de facto* benchmark systems which are frequently used to compare and test different approaches in the field [40, 47, 45, 41, 46, 65, 43, 66].

3.5.1 Benchmarking the self-consistent procedure

In correspondence with an actual calculation using the ATK program (and with the presentation in this chapter), we will benchmark the overall electronic transport modeling as a two-step calculation: First, the self-consistent procedure, which is used to obtain the self-consistent Hamiltonian matrices of electrodes and central region, and second, the subsequent transmission calculations, which is used to obtain the conduction or $I-V$ characteristics etc. In this section, we consider the self-consistent procedure, which is described by the flow diagram of Fig. 2.1 and the discussions in Secs. 3.2.5 and 3.3.5.

In Table 3.1 we display the main data from achieving a self-consistent elec-

Table 3.1: Benchmark results for calculating the self-consistent electronic density with the ATK program for the two-probe systems given in Fig. 3.8. The system type, bias voltage, and the numbers of atoms and matrix sizes N of the central region (electrode unit cell) are indicated. The two right-most columns show the numbers of iterations required and the CPU-times in seconds.

System	Bias	Atoms	N	Iterations	CPU
Al–C×7–Al	-	74(18)	296(72)	16	1065.0
Al–C×7–Al	1V	74(18)	296(72)	17	1572.8
Au–DTB–Au	-	102(3×9)	979(243)	22	7184.2
Au–DTB–Au	1V	102(3×9)	979(243)	20	9461.4
CNT(4,4)–CNT(8,0)	-	256(64)	1024(256)	33	14213.7
CNT(4,4)–CNT(8,0)	1V	256(64)	1024(256)	25	26791.1

tron density in the central region of the two-probe example systems to a tolerance of $||n - n'||_2 < 10^{-4}$Rydberg/Bohr^3. These calculations are carried out for zero bias and 1V bias in order to measure the computational expense in both the equilibrium and non-equilibrium cases, as discussed in Sec. 3.2.5. We note that the results presented are not based on ALGORITHM III in which the full matrix inverse for $\mathbf{G}_C$ is evalutated, but rather the version where only the block tridiagonal part is found, which will be described in Chap. 4. The total number of iterations required for convergence are between 16-33 as indicated. In all cases we use $\mathbf{k}$-point sampling $(1, 1, 100)$ of the Monkhorst type for the bulk electrodes. The last column shows the total CPU-times in seconds. It is apparent that the CPU-times depend strongly on the sizes N of the Hamiltonian matrices at hand, which are proportional to the numbers of atoms present in the C region and the bulk electrodes.[7]

A detailed description of the benchmark results in terms of the percentage of time spent in the individual key parts of the calculation in shown in the pie diagrams in Fig. 3.9. The four most time consuming steps are indicated while the remaining operations are counted together and labeled as the rest. The "Calculate v_{eff}" and "Construct $\mathbf{H}$" steps are well-known from the benchmarking of the molecule and electrode¨systems in Sec. 2.2.2. These parts are quite dominating for the smaller systems Al–C×7–Al and Au–DTB–Au. However, the self-consistent procedure for two-probe systems also consists of the calculation the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$, and the subsequent evaluation of the Green's function matrix $\mathbf{G}_C$, as illustrated in the flow diagram in Fig. 3.5.

[7] Notice that for the Au–DTB–Au system we have reduced the lateral size of the bulk electrode layers to an elementary unit cell of 3 gold atoms, by taking the symmetry properties of the metallic electrode into account. This is a common approach for crystalline electrodes (see, e.g., Ref. [61]) which reduces the computation cost of the self-energy matrices considerably.

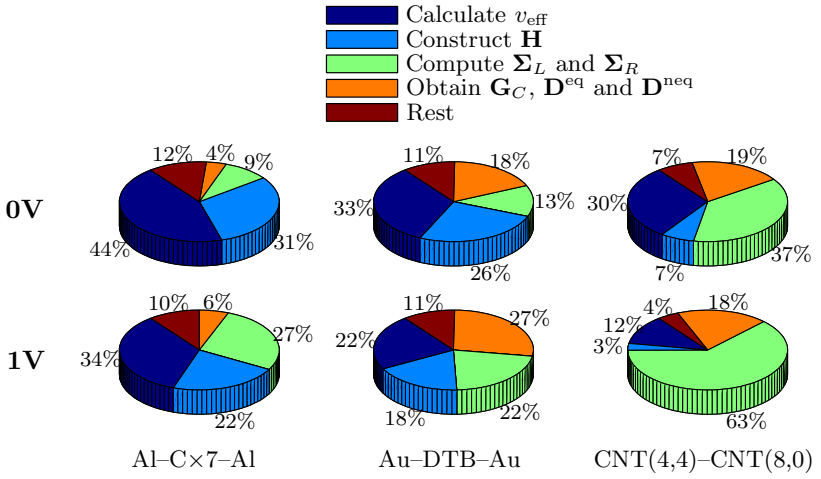


Figure 3.9: Schematic illustration of the percentages of CPU-time spent in the main computational steps of the benchmark calculations listed in Table 3.1.

These tasks are repeated for several energies E in order to obtain the density matrices $\mathbf{D}^{\text{eq}}$ and $\mathbf{D}^{\text{neq}}$ ($\mathbf{D}^{\text{neq}}$ only for the 1V bias cases). We here display the collective time used for computing the self-energy matrices as a distinct task and group the evaluation of $\mathbf{G}_C$,$\mathbf{D}^{\text{eq}}$ and $\mathbf{D}^{\text{neq}}$ into another task.

Fig. 3.9 shows that these latter two tasks, which have been identified and discussed elaborately in this chapter, are increasing rapidly in significance for larger systems. In the case of the final calculation of CNT(4,4)–CNT(8,0) at 1V bias the time spent in these parts of the program corresponds to more than 81% of the overall run-time. For larger systems this percentage will become even higher. It is one of the goals of this work (see Chaps. 4–6) to introduce appropriate optimizations and new methods that can reduce the expense of these computationally most costly parts of electronic transport modeling, in order to be able to investigate larger and more interesting systems in the future.

3.5.2 Benchmarking the transmission calculations

In this section we will benchmark the methods for transmission calculations which we have discussed (i.e., ALGORITHMS III, V and VI). To this end, we determine the transmission coefficient $T(E)$ as a function of the energy E (i.e., the transmission spectrum) for the example systems in Fig. 3.8. It is assumed that the self-consistent Hamiltonian matrices for the example systems are known and stored on disk for easy access. Thus the benchmarks presented in the following correspond only to the second step of the two-step electronic transport modeling calculation.

Table 3.2: CPU-times in seconds for the calculation of the transmission coefficient $T(E)$ at 20 different energies inside $E \in [-2\text{ eV}; 2\text{ eV}]$ for the example systems using the different methods described in this chapter. The time spent in the individual stages of the calculation are explicitly given in columns 2-4.

Al–C×7–Al					
Method	Left	Right	Sweep	Tr{·}	CPU total
Green's	4.5	4.6	0.4	0.0	9.9
WFM	3.7	3.6	0.3	0.1	7.9
Hybrid	3.7	3.6	0.3	0.1	7.9
WFM(Γ-point)	1.8	1.8	0.3	0.1	4.4
Hybrid(Γ-point)	1.9	1.8	0.2	0.1	4.3

Au–DTB–Au					
Method	Left	Right	Sweep	Tr{·}	CPU total
Green's	156.7	155.5	9.6	1.7	337.7
WFM	70.2	67.2	5.2	0.8	158.0
Hybrid	70.7	68.2	5.0	1.7	160.6
WFM(Γ-point)	40.0	39.5	5.1	0.8	100.2
Hybrid(Γ-point)	39.8	38.0	5.1	1.6	98.7

CNT(4,4)–CNT(8,0)					
Method	Left	Right	Sweep	Tr{·}	CPU total
Green's	181.3	147.2	13.8	2.5	348.7
WFM	90.9	153.8	5.8	0.2	254.2
Hybrid	91.3	154.6	5.6	1.9	256.8
WFM(Γ-point)	54.4	65.4	5.8	0.2	129.2
Hybrid(Γ-point)	53.9	64.3	5.5	1.9	128.9

The results from benchmarking the calculation of the transmission spectrum at 20 energy points between -2 eV and 2 eV, relative to the Fermi level of the electrodes, is shown in Table 3.2.[8] In addition to the total CPU time for the overall calculation, we have explicitly measured the time spent in the key stages of the algorithms. These are the following: The evaluation of the self-energy matrices (or equivalent) for the left and right electrodes, respectively, the downwards block Gaussian elimination sweep, and the final transmission calculation using either Caroli's trace expression in Eq. (3.42) or the Landauer-Büttiker formula in Eq. (3.7). In all calculations these four stages represent more than 85% of the total CPU-times which are displayed in the last column.

Comparing the timings for the different methods, we see that the WFM method is in general the fastest approach and more than two times faster than the standard Green's function method for the Au–DTB–Au system. The main reason for this is that the evaluation of the self-energy matrices from solving a QEP (i.e., calling ZGEEV) is much more efficient than from the surface Green's functions obtained with the 2^n recursive algorithm. However, also the other stages are less costly for the WFM method, and this by an increasing factor as the systems grow bigger. It is apparent that the Hybrid method is very similar in efficiency to the WFM method. Again this is because the calculations of the self-energy matrices are the same. The additional time the Hybrid method spends in $\mathrm{Tr}\{\cdot\}$ compared to the standard WFM method is outweighed by less overhead in other parts of the method, which is not explicitly measured.

As an interesting special case we also list benchmark results in Table 3.2 for the corresponding transmission calculations within the Γ-point approximation (see Sec. 2.1.7). This is not a very good approximation in the example systems, but included here for the sake of comparison. The timings indicate the quite significant computational savings that can be achieved with the WFM and Hybrid methods when the Hamiltonian matrices are real as in the case of the Γ-point approximation. Since the Green's function method has a complex infinitesimal part η in the energy, it always requires the use of complex arithmetic when evaluating the self-energy matrices. The WFM and Hybrid method, on the other hand, can use real energy throughout the transmission calculation and therefore solve QEPs in real arithmetic (i.e., by calling DGEEV) in order to obtain the self-energy matrices, which in practice halves the time spent in these stages.

[8] Here we will not interpret or analyze the physics of the transmission spectra obtained. We refer the reader to the references mentioned above for such descriptions.

Chapter 4

Optimizations of the Green's function method

One of the expensive computational task in the Green's function method when used to model electronic transport in two-probe systems is the matrix inverse of the block tridiagonal Hamiltonian matrix. It is unfortunately so, that the inverse of a block tridiagonal matrix fills out, making its straightforward calculation by Block Gaussian elimination an $O(N^2)$ process. This was discussed previously in Sec. 3.2.3. However, if we are only interested in the block tridiagonal part of the inverse of a block tridiagonal matrix, then this can be achieved in $O(N)$ operations. In the first part of this chapter, we will describe a simple and efficient method to do this, based on two independent block Gaussian elimination sweeps. Moreover, in the second part, we will show how these block Gaussian elimination sweeps lead naturally to self-energy matrix definitions for all the layers of the central region, and that this suggests another computationally more efficient calculation of the total transmission $T(E)$. The new algorithms obtained in the two parts of this chapter can be considered optimizations of the Green's function method presented in Sec. 3.2.

4.1 Block tridiagonal matrix inverse

In Sec. 3.2.3 we described an algorithm to calculate the matrix inverse of a block tridiagonal Hamiltonian matrix in order to obtain the Green's function of the central region $\mathbf{G}_C$ in Eq. (3.23). This algorithm had $O(N^2)$ computational complexity for $N \approx \mathrm{n}\bar{m}$, where n is the number layers in the C region, and $\bar{m}$ is the maximum order of any of the Hamiltonian blocks $\bar{\mathbf{H}}_i$, $i = 1, \ldots, \mathrm{n}$.

Let us now discuss how to determine only the block tridiagonal part of $\mathbf{G}_C$ in $O(N)$ operations. From Eqs. (3.36)–(3.38) it seems that in order to calculate for example the diagonal blocks $\mathbf{J}'_{ii}$ of the inverse we have to know $\mathbf{J}_{\mathrm{n},i}$ obtained from applying the full downwards sweep in Eqs. (1.4)–(1.6). This clearly prevents any improvement of the complexity. However, by performing

the downwards and upwards sweeps independently, and both on the original matrix in Eq. (1.3), we can derive an alternative evaluation sequence for the blocks of $\mathbf{G}_C = \mathbf{A}^{-1}$.

If we write down the upwards version of the downwards block Gaussian elimination defined in Eqs. (1.4)–(1.6), we get

$$\sim \left(\begin{array}{cccc|cccc} \mathbf{A}^{\backprime}_{11} & & & & \mathbf{I} & \mathbf{J}_{12} & \dots & \mathbf{J}_{1,\mathrm{n}} \\ \mathbf{A}_{21} & \mathbf{A}^{\backprime}_{22} & & & & \mathbf{I} & \ddots & \vdots \\ & \ddots & \ddots & & & & \ddots & \mathbf{J}_{\mathrm{n}-1,\mathrm{n}} \\ & & \mathbf{A}_{\mathrm{n},\mathrm{n}-1} & \mathbf{A}^{\backprime}_{\mathrm{n},\mathrm{n}} & & & & \mathbf{I} \end{array}\right), \tag{4.1}$$

where $\mathbf{A}^{\backprime}_{\mathrm{n},\mathrm{n}} = \mathbf{A}_{\mathrm{n},\mathrm{n}}$ and

$$\mathbf{A}^{\backprime}_{ii} = \mathbf{A}_{ii} - \mathbf{A}_{i,i+1}(\mathbf{A}^{\backprime}_{i+1,i+1})^{-1}\mathbf{A}_{i+1,i}, \quad i < \mathrm{n} \tag{4.2}$$

$$\mathbf{J}_{ij} = -\mathbf{A}_{i,i+1}(\mathbf{A}^{\backprime}_{i+1,i+1})^{-1}\mathbf{J}_{i+1,j}, \quad i < \mathrm{n}, j < i. \tag{4.3}$$

Furthermore, if we now subtract the downwards result in Eq. (1.4) and the upwards result in Eq. (4.1) from the initial augmented matrix in Eq. (1.3), we arrive at

$$\left(\begin{array}{cccc|cccc} \mathbf{B}_{11} & & & & -\mathbf{I} & -\mathbf{J}_{1,2} & \dots & -\mathbf{J}_{1,\mathrm{n}} \\ & \mathbf{B}_{22} & & & -\mathbf{J}_{2,1} & -\mathbf{I} & \ddots & \vdots \\ & & \ddots & & \vdots & \ddots & \ddots & -\mathbf{J}_{\mathrm{n}-1,\mathrm{n}} \\ & & & \mathbf{B}_{\mathrm{n},\mathrm{n}} & -\mathbf{J}_{\mathrm{n},1} & \cdots & -\mathbf{J}_{\mathrm{n},\mathrm{n}-1} & -\mathbf{I} \end{array}\right), \tag{4.4}$$

where the diagonal blocks of the left-hand side matrix are

$$\mathbf{B}_{ii} = \mathbf{A}_{ii} - \mathbf{A}'_{ii} - \mathbf{A}^{\backprime}_{ii}, \quad 1 \le i \le \mathrm{n}. \tag{4.5}$$

Then, because Eq. (4.4) is equal to $(\mathbf{A} - \mathbf{A} - \mathbf{A}|\mathbf{I} - \mathbf{I} - \mathbf{I}) = -(\mathbf{A}|\mathbf{I})$, we can make LU-factorizations of the diagonal blocks $\mathbf{B}_{ii}$ and multiply $(\mathbf{B}_{ii})^{-1}$ onto the ith row of the augmented matrix in Eq. (4.4), which results in the identity matrix in the left-hand side and $-\mathbf{A}^{-1}$ in the right-hand side. By inspection one can realize that to calculate the diagonal blocks $[\mathbf{A}^{-1}]_{ii}$ now require $3\mathrm{n}-2$ LU-factorizations of the matrices $\mathbf{A}'_{ii}$ $(i = 2, \dots, \mathrm{n})$, $\mathbf{A}^{\backprime}_{ii}$ $(i = 1, \dots, \mathrm{n}-1)$, and $\mathbf{B}_{ii}$ $(i = 1, \dots, \mathrm{n})$, and only $5\mathrm{n}-4$ multiplications and $4\mathrm{n}-6$ additions (none is necessary for $\mathbf{B}_{11}$ and $\mathbf{B}_{\mathrm{n},\mathrm{n}}$). Moreover, in order to obtain the off-diagonal blocks $[\mathbf{A}^{-1}]_{i,i\pm1}$ of the inverse matrix, only a single additional multiplication per block is needed.

The linear transformations in this alternative method for the matrix inverse outline the steps of a simple sequential algorithm to obtain the block tridiagonal part of $\mathbf{G}_C$, based on two (independent) block Gaussian elimination sweeps

and the inversion (i.e., LU-factorization) of the diagonal blocks $\mathbf{B}_{ii}$. We will implement it as follows,

ALGORITHM VIII: *Block tridiagonal Green's function matrix*

1. initialize $\mathbf{A}'_1 := \bar{\mathbf{H}}_{11} - \mathbf{\Sigma}_L,\ \mathbf{A}^{\backslash}_{\mathrm{n}} := \bar{\mathbf{H}}_{\mathrm{n,n}} - \mathbf{\Sigma}_R,\ \mathbf{B} := \mathbf{I}$
2. **for** $i := 2, \ldots, \mathrm{n}$
3. solve $\bar{\mathbf{H}}_{i,i-1} = \mathbf{J}'_i \mathbf{A}'_{i-1}$ for $\mathbf{J}'_i$
4. $\mathbf{A}'_i := \bar{\mathbf{H}}_{i,i} - \mathbf{J}'_i \bar{\mathbf{H}}_{i-1,i}$
5. **end**

downwards sweep

6. **for** $i := \mathrm{n}-1, \ldots, 1$
7. solve $\bar{\mathbf{H}}_{i,i+1} = \mathbf{J}^{\backslash}_i \mathbf{A}^{\backslash}_{i+1}$ for $\mathbf{J}^{\backslash}_i$
8. $\mathbf{A}^{\backslash}_i := \bar{\mathbf{H}}_{i,i} - \mathbf{J}^{\backslash}_i \bar{\mathbf{H}}_{i+1,i}$
9. **end**

upwards sweep

(4.6)

10. $\mathbf{B}_1 := -\mathbf{A}^{\backslash}_1 + \mathbf{\Sigma}_L,\ \mathbf{B}_{\mathrm{n}} := -\mathbf{A}'_{\mathrm{n}} + \mathbf{\Sigma}_R$
11. solve $\mathbf{B}_1[\mathbf{G}_{11}\ \mathbf{G}_{12}] = [\mathbf{I}\ \mathbf{J}^{\backslash}_1]$
12. solve $\mathbf{B}_{\mathrm{n}}[\mathbf{G}_{\mathrm{n,n-1}}\ \mathbf{G}_{\mathrm{n,n}}] = [\mathbf{J}'_{\mathrm{n}}\ \mathbf{I}]$
13. **for** $i := 2, \ldots, \mathrm{n}-1$
14. $\mathbf{B}_i := \bar{\mathbf{H}}_{i,i} - \mathbf{A}'_i - \mathbf{A}^{\backslash}_i$
15. solve $\mathbf{B}_i[\mathbf{G}_{i,i-1}\ \mathbf{G}_{ii}\ \mathbf{G}_{i,i+1}] = [\mathbf{J}'_i\ \mathbf{I}\ \mathbf{J}^{\backslash}_i]$
16. **end**

which is clearly an $O(N)$ method. Notice that since the two sweeps and all iterations of the final loop are completely independent calculations, the algorithm in Eq. (4.6) is well suited for parallel execution on a dual-core CPU or two CPUs on a parallel computer. Further details and analysis of the above procedure can be found in **PAPER I**.

4.1.1 Basic operations count

As mentioned in Sec. 3.5.1, the better performance of the $O(N)$ method in ALGORITHM VIII , compared to the performance of the $O(N^2)$ method in ALGORITHM III, was already taken into account in the benchmarking results of that section. The reason for this is, that the above algorithm was adopted in the commercial software that contains the ATK program [23]. In fact, the version of ATK which is used as a baseline implementation in this thesis (version 2.0) was released after ALGORITHM VIII had been incorporated as the standard routine for obtaining $\mathbf{G}_C$ in the self-consistent DFT procedure. Therefore, we do not find it necessary to explicitly measure CPU-times in order to demonstrate its applicability. The computational savings are indeed significant and general as can be realized from simply counting the basic matrix operations, see Table 4.1.

Table 4.1: The number of basis block operations performed using ALGORITHM III to compute the full matrix inverse for $\mathbf{G}_C$, and using ALGORITHM VIII to compute the block tridiagonal part of of the inverse for $\mathbf{G}_C$.

Algorithm	LU-factorizations	Multiplications	Additions
ALGORITHM III	$3n-2$	n^2+4n-4	$4n-6$
ALGORITHM VIII	$3n-2$	$7n-6$	$4n-6$

4.2 Efficient transmission calculations

As we have seen in Sec. 3.2, the Green's function block $[\mathbf{G}_C]_{n,1}$ of the central region contains the information needed to determine the total transmission $T(E)$ when the coupling to the electrodes via the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ are known. However, any other block (i,j) of the Green's function $[\mathbf{G}_C]_{i,j}$ could in principle be used if the corresponding self-energy matrices were available. For the block $[\mathbf{G}_C]_{i,j}$ we would need the self-energy matrices corresponding to the the case where the ith and jth principal layers are the border layers of the central region. We will now describe an intuitive scheme of extending the self-energy concept to all layers of the central region in order to use the smallest block on the diagonal of $\mathbf{G}_C$ in the usual Green's function transmission expression in Eq. (3.42). The sheme is simple, computationally advantageous, and gives a more accurate result than using the standard approach described in Sec. 3.2.4.

4.2.1 Generalized self-energy matrices

Our first step is to derive a matrix expression that can be interpreted as the self-energy for a given layer of the C region by means of purely algebraic manipulations. We will use the general inversion result for 2×2 block matrices in Eq. (1.2). Consider the L–C part of the two-probe Hamiltonian $\bar{\mathbf{H}}$ in Eq. (3.11). Let us associate $\mathbf{A}_{11}$ with the semi-infinite submatrix of $\bar{\mathbf{H}}$ and $\mathbf{A}_{22}$ with the finite central region matrix $\bar{\mathbf{H}}_C$, such that

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \left(\begin{array}{cc|c} \ddots & \ddots & \\ \ddots & \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} \\ & \bar{\mathbf{H}}_{L,L}^{\dagger} & \begin{pmatrix} & \\ & \bar{\mathbf{H}}_C \end{pmatrix} \end{array}\right). \tag{4.7}$$

We can now apply the inversion formula in Eq. (1.2) in order to obtain the inverse of the $\mathbf{A}_{22}$ block and in the process recognize the second term of the Schur complement $\mathbf{S} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ as the usual self-energy matrix $\mathbf{\Sigma}_L =$

$\bar{\mathbf{H}}_{L,L}^{\dagger} \boldsymbol{g}_L \bar{\mathbf{H}}_{L,L}$ defined in Sec. 3.2.1, where $\boldsymbol{g}_L$ is the surface Green's function of the left electrode. Let us then change notation by adding a subscript 1 to denote the layer where this self-energy matrix is defined (i.e., it is non-zero only in block (1,1)), and also making the subscript L into a superscript, i.e., $\boldsymbol{\Sigma}_1^L \equiv \boldsymbol{\Sigma}_L$.

Suppose now that the $\mathbf{A}_{11}$ block is augmented to include the first block column and top block row of $\bar{\mathbf{H}}_C$ (whose sizes are equal to those of $\bar{\mathbf{H}}_{LL}$), and that $\mathbf{A}_{22}$ is correspondingly stripped. Again we can attempt to obtain the inverse of the smaller $(2,2)$-block and consider the second term of the corresponding Schur complement, which now takes the form $\mathbf{S} = \bar{\mathbf{H}}_1 - \boldsymbol{\Sigma}_2^L$, where $\boldsymbol{\Sigma}_2^L = \bar{\mathbf{H}}_{1,2}^{\dagger}(\bar{\mathbf{H}}_1 - \boldsymbol{\Sigma}_1^L)^{-1}\bar{\mathbf{H}}_{1,2}$. We will interpret matrix $\boldsymbol{\Sigma}_2^L$ as a self-energy matrix for the left electrode plus the left-most layer of the central region.

Since the system is block tridiagonal, this procedure can be continued for $i = 2, 3, \ldots, \mathrm{n}$, which corresponds to performing block Gaussian elimination of the lower band. For each elimination step we can interpret the corresponding matrix $\boldsymbol{\Sigma}_i^L$ as the self-energy matrix describing the coupling of the left electrode augmented with the i leftmost layers of the central region. A recursion expression to obtain these matrices is then given by

$$\boldsymbol{\Sigma}_i^L = \bar{\mathbf{H}}_{i-1,i}^{\dagger}(\bar{\mathbf{H}}_{i-1} - \boldsymbol{\Sigma}_{i-1}^L)^{-1}\bar{\mathbf{H}}_{i-1,i}, \quad -\infty < i \leq k, \tag{4.8}$$

where k is a given layer in the central region.

In the same manner, self-energy matrices $\boldsymbol{\Sigma}_i^R$ can be defined that represent the coupling of the right semi-infinite part in Eq. (3.11) to the central region and the successive extension of the right region by one block column and block row of $\bar{\mathbf{H}}_C$. This corresponds to performing a block Gaussian elimination of the upper band and we define the $\boldsymbol{\Sigma}_i^R$ matrices by writing

$$\boldsymbol{\Sigma}_i^R = \bar{\mathbf{H}}_{i,i+1}(\bar{\mathbf{H}}_{i+1} - \boldsymbol{\Sigma}_{i+1}^R)^{-1}\bar{\mathbf{H}}_{i,i+1}^{\dagger}, \quad k \leq i < \infty. \tag{4.9}$$

as the corresponding recursion expression.

Eqs. (4.8) and (4.9) describe the successive coupling of adjacent layers from within the left and right electrodes to layer k within the central region. This makes it straightforward to recursively evaluate the self-energy matrices $\boldsymbol{\Sigma}_k^L$ and $\boldsymbol{\Sigma}_k^R$ for a given layer k since the matrices $\boldsymbol{\Sigma}_1^L$ and $\boldsymbol{\Sigma}_{\mathrm{n}}^R$ are available from either the Green's function method or the hybrid method (i.e., $\boldsymbol{\Sigma}_1^L = \boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_{\mathrm{n}}^R = \boldsymbol{\Sigma}_R$). We will use this to speed up the transmission calculations.

4.2.2 Fast transmission calculations

Consider the calculation of the total transmission probability $T(E)$ for a given energy E. As described in Sec. 3.2.4, the most common way to proceed in the Green's function approach when $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_R$ have been obtained is to determine the off-diagonal block $[\boldsymbol{G}_C]_{\mathrm{n},1}$ and use this in the Caroli formula in Eq. (3.42). This is for example the case in Refs. [4, 43, 39, 37]. Here we suggest a different approach, in which the self-energy matrices are propagated from their corner

block positions further inwards in the central region Hamiltonian. We note that in order to obtain the information required to calculate the transmission through the entire two-probe system, it is necessary to determine the coupling between all layers connecting the left with the right electrodes. This can be initiated from left to right, from right to left, or from both ends.

The idea is then to select a block $\bar{\mathbf{H}}_k$ on the diagonal of $\bar{\mathbf{H}}_C$, typically the block of smallest size, and attempt the compute block $[\mathbf{G}_C]_{k,k}$ of the corresponding Green's function matrix. Clearly, this block also contains the information needed to determine $T(E)$ if the coupling to layer $k-1$ and layer $k+1$ is known. We therefore calculate the self-energy matrices $\mathbf{\Sigma}_i^L$ $(i = 1, \ldots, k)$ and $\mathbf{\Sigma}_i^R$ $(i = k, \ldots, \mathrm{n})$, describing the coupling from the left and right electrodes to layer k using the recursion expressions Eq. (4.8) and Eq. (4.9). Subsequently, this allows us to rewrite the original equation $(\bar{\mathbf{H}}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R)\mathbf{G}_C = \mathbf{I}$ defining the Green's function of the finite central region, as

$$\left(\begin{array}{c|c|c} \left(\begin{array}{c} \bar{\mathbf{H}}_{\downarrow}^{L} \end{array}\right) & \bar{\mathbf{H}}_{k-1,k} & \\ \hline & \mathbf{H}_k - \mathbf{\Sigma}_k^L - \mathbf{\Sigma}_k^R & \\ \hline & \bar{\mathbf{H}}_{k,k+1}^{\dagger} & \left(\begin{array}{c} \bar{\mathbf{H}}_{\uparrow}^{R} \end{array}\right) \end{array}\right) \times \mathbf{G}_C =$$

$$\left(\begin{array}{cccc|c|cccc} \mathbf{I} & & & & & & & & \\ \mathbf{J}_2^L & \mathbf{I} & & & & & & & \\ \mathbf{J}_3^L\mathbf{J}_2^L & \mathbf{J}_3^L & \ddots & & & & & & \\ \vdots & \vdots & \ddots & \mathbf{I} & & & & & \\ \hline \mathbf{J}_k^L\ldots\mathbf{J}_2^L & \mathbf{J}_k^L\ldots\mathbf{J}_3^L & \cdots & \mathbf{J}_k^L & \mathbf{I} & \mathbf{J}_k^R & \cdots & \mathbf{J}_k^R\ldots\mathbf{J}_{\mathrm{n}-2}^R & \mathbf{J}_k^R\ldots\mathbf{J}_{\mathrm{n}-1}^R \\ \hline & & & & & \mathbf{I} & \ddots & \vdots & \vdots \\ & & & & & & \ddots & \mathbf{J}_{\mathrm{n}-2}^R & \mathbf{J}_{\mathrm{n}-2}^R\mathbf{J}_{\mathrm{n}-1}^R \\ & & & & & & & \mathbf{I} & \mathbf{J}_{\mathrm{n}-1}^R \\ & & & & & & & & \mathbf{I} \end{array}\right), \tag{4.10}$$

where $\bar{\mathbf{H}}_{\downarrow}^{L}$ is the upper-bidiagonal matrix

$$\bar{\mathbf{H}}_{\downarrow}^{L} = \begin{pmatrix} \bar{\mathbf{H}}_1 - \mathbf{\Sigma}_1^L & \bar{\mathbf{H}}_{1,2} & & \\ & \ddots & \ddots & \\ & & \ddots & \bar{\mathbf{H}}_{k-2,k-1} \\ & & & \bar{\mathbf{H}}_{k-1} - \mathbf{\Sigma}_{k-1}^L \end{pmatrix}, \tag{4.11}$$

the $\bar{\mathbf{H}}_{\uparrow}^{L}$ is the lower-bidiagonal matrix

$$\bar{\mathbf{H}}_{\uparrow}^{R} = \begin{pmatrix} \bar{\mathbf{H}}_{k+1} - \mathbf{\Sigma}_{k+1}^{R} & & & \\ \bar{\mathbf{H}}_{k+1,k+2}^{\dagger} & \ddots & & \\ & \ddots & \ddots & \\ & & \bar{\mathbf{H}}_{\mathrm{n}-1,\mathrm{n}}^{\dagger} & \bar{\mathbf{H}}_{\mathrm{n}} - \mathbf{\Sigma}_{\mathrm{n}}^{R} \end{pmatrix}, \tag{4.12}$$

and the off-diagonal blocks appearing in the right-hand side identity matrix, which is a result of the block Gaussian elimination steps, are given by

$$\mathbf{J}_{i}^{L} = \bar{\mathbf{H}}_{i-1,i}^{\dagger}(\bar{\mathbf{H}}_{i-1} - \mathbf{\Sigma}_{i-1}^{L})^{-1}, \quad -\infty < i \leq k \tag{4.13}$$

$$\mathbf{J}_{i}^{R} = \bar{\mathbf{H}}_{i,i+1}(\bar{\mathbf{H}}_{i+1} - \mathbf{\Sigma}_{i+1}^{R})^{-1}, \quad k \leq i < \infty. \tag{4.14}$$

Inspecting Eq. (4.10), we notice that the kth row of the left-hand side matrix and the kth column of the right-hand side matrix are non-zero only in the center block. It is therefore simple to obtain the the kth diagonal block of the Green's function,

$$[\mathbf{G}_C]_{k,k} = (\bar{\mathbf{H}}_k - \mathbf{\Sigma}_k^L - \mathbf{\Sigma}_k^R)^{-1}, \tag{4.15}$$

which corresponds to inverting the block of smallest size in the system, if k is chosen accordingly.

The motivation for performing the above calculations is to determine the electron transmission through the full block tridiagonal system. By applying Eq. (4.8) and Eq. (4.9) to evaluate the self-energy matrices $\mathbf{\Sigma}_k^L$ and $\mathbf{\Sigma}_k^R$ and subsequently Eq. (4.15) to obtain $[\mathbf{G}_C]_{k,k}$, the transmission $T(E)$ can be computed from the appropriate Caroli formula

$$T(E) = \mathrm{Tr}\{\mathbf{\Gamma}_k^L \mathbf{G}_{k,k}^{\dagger} \mathbf{\Gamma}_k^R \mathbf{G}_{k,k}\}, \tag{4.16}$$

where we have used $\mathbf{G}$, corresponding to the full system Green's function, instead of $\mathbf{G}_C$, since this represents the same block in our setup, and

$$\mathbf{\Gamma}_k^L = i(\mathbf{\Sigma}_k^L - \mathbf{\Sigma}_k^{L\dagger}), \quad \mathbf{\Gamma}_k^R = i(\mathbf{\Sigma}_k^R - \mathbf{\Sigma}_k^{R\dagger}), \tag{4.17}$$

are the broadening matrices defined for the layer k. Eq. (4.16) is valid due to the properties of the trace operator and the specific structure of Eq. (4.10) (the explicit proof is given in **PAPER I**).

Our implementation of this fast transmission Green's function method can

be written:

ALGORITHM IX: *Calculate* $T(E)$ *using* $\mathbf{G}_{k,k}$

1. initialize $\mathbf{\Sigma}_1^L := \mathbf{\Sigma}_L,\ \mathbf{\Sigma}_{\mathrm{n}}^R := \mathbf{\Sigma}_R$
2. **for** $i := 2, \ldots, k$
3. solve $(\bar{\mathbf{H}}_{i-1} - \mathbf{\Sigma}_{i-1}^L)\mathbf{X} = \bar{\mathbf{H}}_{i-1,i}$
4. $\mathbf{\Sigma}_i^L := \bar{\mathbf{H}}_{i-1,i}^\dagger \mathbf{X}$
5. **end**
6. **for** $i := \mathrm{n}-1, \ldots, k$
7. solve $(\bar{\mathbf{H}}_{i+1} - \mathbf{\Sigma}_{i+1}^R)\mathbf{X} = \bar{\mathbf{H}}_{i,i+1}^\dagger$
8. $\mathbf{\Sigma}_i^R := \bar{\mathbf{H}}_{i,i+1}\mathbf{X}$
9. **end**
10. solve $(\bar{\mathbf{H}}_k - \mathbf{\Sigma}_k^L - \mathbf{\Sigma}_k^R)\mathbf{G}_{k,k} = \mathbf{I}$
11. obtain $T(E)$ from Eqs. (4.16) and (4.17)

(4.18)

In this implementation, all that is needed to calculate $T(E)$ is the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ to initiate the recursion (their evaluation is the focus of the Krylov subspace algorithm presented in Chap. 6) and, after the recursion, the inversion of a single block and a few matrix-matrix multiplications (we can neglect the additions of complexity $O(\bar{m}^2)$). This approach is therefore very efficient. In particular, compared to the case where the off-diagonal block $\mathbf{G}_{\mathrm{n},1}$ is used in the Caroli expression, one saves significantly in floating point operations during the sweep process since the number of explicit matrix-matrix multiplication performed is exactly half (cf. ALGORITHM IV). Also the order of the matrices $\mathbf{G}_{k,k}$,$\mathbf{\Sigma}_k^L$,and $\mathbf{\Sigma}_k^R$, used in the last steps of ALGORITHM IX is often smaller than the corresponding matrices $\mathbf{G}_{\mathrm{n},1}$,$\mathbf{\Sigma}_L$,and $\mathbf{\Sigma}_R$ used in ALGORITHM IV (see more comparison details in **PAPER I**).

4.2.3 Benchmarking the new algorithm

In order to demonstrate the speed-up achieved by the new transmission algorithm, we apply it to the example two-probe systems introduced in Sec. 3.5 and compare with the standard Green's function method. The benchmarking results are displayed in Table 4.2 and show clearly the savings during the sweep stage of the calculation when using the new approach. As explained, this is due to the fewer matrix-matrix multiplication necessary in order to obtain $\mathbf{G}_{k,k}$ instead of $\mathbf{G}_{\mathrm{n},1}$. For the smallest system (Al–C×7–Al) most of the gain disappears in computational overhead. In the case of the two largest systems (Au–DTB–Au, CNT(4,4)–CNT(8,0)), the savings are close to the estimated factor of 2.

We note, however, that the overall speed-up is not very significant because the evaluation of the self-energy matrices, in these two-probe systems, are much

Table 4.2: CPU-times in seconds for the calculation of the transmission coefficient $T(E)$ at 20 different energies inside $E \in [-2\text{ eV}; 2\text{ eV}]$ for the example systems in Fig. 3.8 from applying the standard Green's function method (ALGORITHM IV) and the new fast method (ALGORITHM IX). The time spent in the individual stages of the calculation are explicitly given in columns 2-4.

System	Method	Left	Right	Sweep	Tr{·}	Total
Al–C×7–Al	Green's	4.5	4.6	0.4	0.0	9.9
Al–C×7–Al	Fast Green's	4.5	4.5	0.3	0.0	9.7
Au–DTB–Au	Green's	156.7	155.5	9.6	1.7	337.7
Au–DTB–Au	Fast Green's	157.4	155.9	5.0	1.6	334.1
CNT(4,4)–CNT(8,0)	Green's	181.3	147.2	13.8	2.5	348.7
CNT(4,4)–CNT(8,0)	Fast Green's	181.7	146.9	7.1	2.5	342.0

more expensive than the sweep stage of the calculation. For other systems, in which the central region is relatively larger (see, e.g., Sec. 5.7), the impact of the savings in the sweep stage is correspondingly bigger.

Chapter 5

Efficient wave function matching method

The wave function matching (WFM) method and similar techniques have recently been successfully applied for the calculation of electronic transport in quantum two-probe systems [39, 40, 59, 55, 33, 60, 52, 61, 62]. We have described the details of this method in Sec. 3.3 and implemented the corresponding algorithms (ALGORITHM V/VI). In terms of efficiency they are comparable to the widely used Green's function approach. To our knowledge, the WFM schemes presented so far in the literature requires the evaluation of all the Bloch and evanescent bulk modes of the left and right electrodes in order to obtain the correct coupling between device and electrode regions. The reason for this is that it requires the complete set of bulk modes to be able to represent the proper reflected and transmitted wave functions. In this chapter we will describe a new modified WFM approach that allows for the exclusion of the vast majority of the evanescent modes in all parts of the calculation by simply extending the central region with a few layers. This approach makes it feasible to apply iterative techniques (e.g, as described in the next chapter) to efficiently determine the relatively few bulk modes of interest, which allows for a significant reduction of the computational expense of the WFM method in practice.

5.1 Introduction and motivation

The WFM method is based upon direct matching of the bulk modes in the left and right electrode to the scattering wave function of the central region. For the most part this involves two major tasks; obtaining the bulk electrode modes and solving a system of linear equations. For the purpose of calculating transport properties the application of the method results in the transmission and reflection matrices $\mathbf{t}$ and $\mathbf{r}$ from Eqs. (3.84) and (3.85), respectively. The elements of $\mathbf{t}$ and $\mathbf{r}$ are then used in order to determine the total transmission $T(E)$ or reflection $R(E)$ (where $T+R=1$) within the Landauer-Büttiker theory.

Table 5.1: CPU-times in seconds when using the standard WFM method in ALGORITHM V for calculating $\boldsymbol{t}$ and $T(E)$ at 20 different energies inside $E \in [-2\text{ eV}; 2\text{ eV}]$ for various two-probe systems. The numbers of atoms in the central region (electrode unit cell) are indicated. The two right-most columns show the percentage of the CPU-time used for computing the electrode bulk modes with DGEEV vs. solving the central region linear systems in Eq. (3.91).

System	Atoms	CPU	DGEEV	Eq. (3.91)
Li–Li	32(8)	0.1	75%	25%
Fe–MgO–Fe	27(6)	2.2	61%	38%
Al–C×7–Al	74(18)	4.2	87%	10%
Au–DTB–Au	102(27)	100.2	90%	8%
Au–CNT(8,0)×1–Au	140(27)	90.0	86%	12%
Au–CNT(8,0)×5–Au	268(27)	130.0	65%	34%
CNT(8,0)–CNT(8,0)	192(64)	136.4	94%	5%
CNT(4,4)–CNT(8,0)	256(64|64)	129.2	94%	5%
CNT(5,0)–CNT(10,0)	300(40|80)	141.0	80%	17%
CNT(18,0)–CNT(18,0)	576(144)	1565.7	87%	11%

In practice, it turns out that for many average size two-probe systems, the most time consuming stage of the WFM method is to determine the electrode bulk modes, which requires solving the quadratic eigenvalue problem in Eq. (3.56) for both electrodes. As examples, see the profiling results listed in Table 5.1, where we have used the method to compute $\boldsymbol{t}$ and $T(E)$ for a selection of two-probe systems (timings are only for the final transmission calculations, not for obtaining the self-consistent Hamiltonians). We should point out that the metallic electrodes in the two-probe systems considered in Table 5.1 can be fully described by much smaller unit cells than indicated (often only a few atoms are needed) and therefore the time spend on computing the bulk modes can be vastly reduced in these specific cases. For a general method, however, which supports CNTs, nano wires, etc., as the electrodes, the measurements are appropriate for showing the overall trend in the computational costs.

The results in Table 5.1 show that to determine the bulk modes by employing the state-of-the-art LAPACK eigensolver DGEEV [29] is, in general, much more expensive than to solve the system of linear equations in Eq. (3.71). We expect this trend to hold for larger systems as well. Therefore, in the attempt to model significantly larger devices (thousands of atoms), it is of essential interest to reduce the numerical cost of the bulk modes calculation. We will argue that a physically reasonable and computationally attractive approach is to limit the number of bulk modes taken into account, e.g., by excluding the least important evanescent modes. In the remainder of this chapter, a new technique to do this

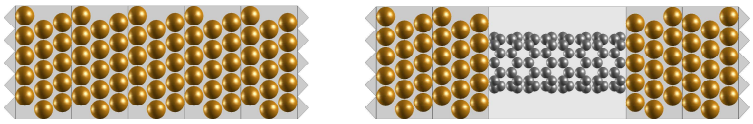


Figure 5.1: Atomic configuration of gold electrode example systems. Left: the ideal Au(111) electrode. Right: the Au–CNT(8,0)×4–Au two-probe system.

in a rigorous and systematic fashion is presented.

Before we begin let us stress that the outset of the following scheme is the WFM method in the "combined" formulation derived in Sec. 3.4, and not the original formulation by Refs. [52, 55, 62] which was presented in Sec. 3.3. The analysis and conclusions hold, however, for both formulations. For notational simplicity in the following sections, we leave out the implied subscripts L or R, indicating the left or right electrode, whenever the formalism is the same for both (e.g, for symbols $m, \lambda_k, \mathbf{c}_k, \mathbf{C}^{\pm}, \mathbf{\Lambda}^{\pm}, \mathbf{B}^{\pm}, \mathbf{\Sigma}$, etc.).

5.2 Decay of evanescent bulk modes

As discussed in Sec. 3.3.2, an electron in a periodic bulk system can be represented by a Bloch wave (e.g., mode $\mathbf{c}_k$ of the electrode) in agreement with Bloch's theorem. A corollary of this result is that the Bloch wave vector $\mathbf{k}$ is a conserved quantity in a bulk system (modulo addition of reciprocal lattice vectors), and hence that the group velocity v_k of the mode is conserved. In other words, the electrons can propagate without scattering through a bulk electrode, almost like free particles, and any electrical resistance in such a conductor is only a result from things like imperfections that break the periodicity.

In the two-probe systems considered in this work the electrodes are assumed to be ideal and periodicity is only broken because of the device in the central region. It commends an intuitive description in terms of scattering waves as presented in Sec. 3.3.1, for which the incoming, reflected and transmitted wave functions are represented in the (complete) basis of bulk modes. The procedure to determine the Bloch factors λ_k and non-trivial modes $\mathbf{c}_k$ of an ideal electrode and subsequently characterize these as right-going (+) or left-going (−) was given in Sec. 3.3.2. We note that only the obtained propagating modes with $|\lambda_k| = 1$ are able to carry charge deeply into the electrodes and thus enter in the Landauer expression in Eq. (3.8). The evanescent modes with $|\lambda_k| \neq 1$, on the other hand, decay exponentially but can still contribute to the current in the two-probe system, as the "tails" may reach across the central region boundaries.

Consider a typical example of an electrode mode evaluation: We look at the gold electrode in the left part of Fig. 5.1 with 27 atoms in the unit cell represented by 9 (sp^3d^5) orbitals for each Au-atom. Such a system results

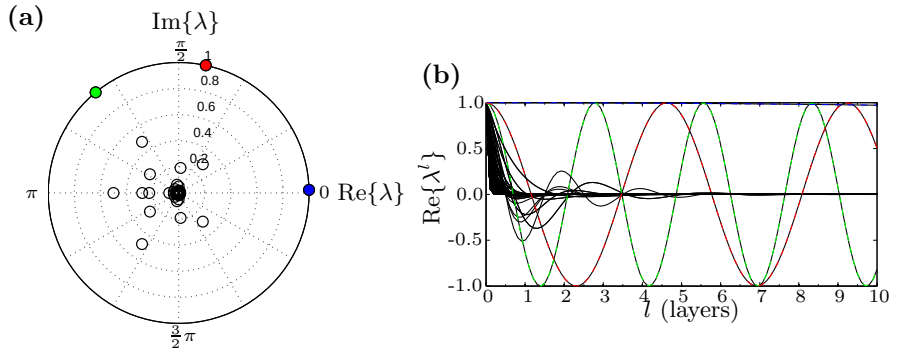


Figure 5.2: (a) Positions of the Bloch factors λ_k ($|\lambda_k| \leq 1$) obtained for a bulk Au(111) electrode with 27 atoms per unit cell at $E = -1.5$ eV. (b) Amplitudes of the corresponding normalized electrode modes $\mathbf{c}_k$ moving through 10 principal layers of the ideal bulk electrode. A total of 243 modes are shown of which 3 are propagating (colored/dashed) and the rest are evanescent (circles/black).

in 243 right-going and 243 left-going modes. Fig. 5.2a shows the positions in the complex plane of the Bloch factors corresponding the right-going modes (i.e., $|\lambda_k| \leq 1$) for energy $E = -1.5$ eV. We see that there are exactly three propagating modes, which have Bloch factors located on the unit circle. The remaining modes are evanescent, of which many have Bloch factors with small magnitude very close to the origin.

Fig. 5.2b illustrates how the 243 left-going modes would propagate through 10 successive gold electrode principal layers. The figure shows that the amplitudes of the three propagating modes are unchanged, while the evanescent modes are decaying exponentially. In particular, we note that the evanescent modes with Bloch factors of small magnitude are very rapidly decaying and vanishes in comparison to the propagating modes after only a few layers. In the following, we will exploit this observation and attempt to exclude such evanescent modes from the WFM calculation altogether.

5.3 Excluding evanescent modes

In practice we will exclude the rapidly decaying evanescent modes by employing the selection criteria of Khomyakov *et al.* [55] which was given in Eq. (3.60) as part of the classification procedure. More specifically, only the bulk modes $\mathbf{c}_k$ with Bloch factors λ_k satisfying

$$\lambda_{\min} \leq |\lambda_k| \leq \lambda_{\min}^{-1}, \tag{5.1}$$

are computed and subsequently taken into account, however, for a setting of the parameter $\lambda_{\min}$ ($0 < \lambda_{\min} < 1$) that is much higher than in the case of the standard WFM method. Eq. (5.1) is thus adopted as the key relation to select a particular subset of the available bulk modes.

Suppose that we have determined all the non-trivial solutions $(\mathbf{c}_k, \lambda_k)$ of the bulk electrode QEP (Eq. (3.56)) and stored these in the matrices $\mathbf{C}^{\pm}$ and $\mathbf{\Lambda}^{\pm}$ (as noted above we leave out the the appropriate subscripts L and R). We will then denote the mode and Bloch factor matrices from which the rapidly decaying evanescent modes are excluded via Eq. (5.1), and also the Bloch matrices and self-energy matrices obtained from these, with a tilde, i.e., as $\tilde{\mathbf{C}}^{\pm}$,$\tilde{\mathbf{\Lambda}}^{\pm}$, $\tilde{\mathbf{B}}^{\pm}$ and $\tilde{\mathbf{\Sigma}}$. The mode matrices holding the excluded modes are subsequently denoted by a math-ring accent $\mathring{\mathbf{C}}^{\pm}$, so that

$$\mathbf{C}^{\pm} = [\tilde{\mathbf{C}}^{\pm}, \mathring{\mathbf{C}}^{\pm}], \tag{5.2}$$

is the assumed splitting of the full set of modes. All expressions to evaluate the Bloch and self-energy matrices are unchanged as given in Sec. 3.3.2. However, since the column spaces of $\tilde{\mathbf{C}}^{\pm}$ are not complete, there is no longer any guaranty that WFM can be performed so that the resulting self-energy matrices and, in turn, the solution $\mathbf{c}_C = [\mathbf{c}_1^{\mathrm{T}}, \ldots, \mathbf{c}_{\mathrm{n}}^{\mathrm{T}}]^{\mathrm{T}}$ of the linear system in Eq. (3.91), are correct. In addition, errors can occur in the calculation of $\mathbf{t}$ and $\mathbf{r}$ from Eqs. (3.84) and (3.85) because the boundary wave functions $\mathbf{c}_{\mathrm{n}+1}$ and $\mathbf{c}_0$ might not be fully represented in the reduced sets $\tilde{\mathbf{C}}_R^{+}$ an $\tilde{\mathbf{C}}_L^{-}$, respectively.

As shown in Sec. 3.3.3, the key to deriving the WFM block tridiagonal linear system in Eq. (3.91) is twofold: Specifying the layer wave functions coming into the C region and matching the layer wave functions across the C region boundaries. In our case the incoming waves are

$$\mathbf{c}_1^{+} = \mathbf{C}_L^{+} \Lambda_L^{+} \mathbf{a}^{\mathrm{in}}, \tag{5.3}$$

from the left (cf. Eq. (3.64)) and

$$\mathbf{c}_n^{-} = \mathbf{0}, \tag{5.4}$$

from the right. The matching is accomplished by using the Bloch matrices $\mathbf{B}^{\pm} = \mathbf{C}^{\pm}\mathbf{\Lambda}^{\pm}(\mathbf{C}^{\pm})^{-1}$, which by construction propagate the layer wave functions in the bulk electrode, i.e.,

$$\mathbf{c}_j^{\pm} = (\mathbf{B}^{\pm})^{j-i}\mathbf{c}_i^{\pm}, \tag{5.5}$$

where subscript L is implied for the left electrode ($i, j \leq 1$), and R for the right electrode ($i, j \geq n$). Notice that the Bloch matrices are always square and also invertible since any trivial QEP solutions ($\lambda_k \sim 0, \texttt{inf}$) are rejected from the outset in the mode classification procedure, see Sec. 3.3.2. When the reduced Bloch matrices $\tilde{\mathbf{B}}^{\pm}$ are used instead of $\mathbf{B}^{\pm}$, however, the possible components of the wave functions $\mathbf{c}_1$ and $\mathbf{c}_{\mathrm{n}}$ outside the column spaces of $\tilde{\mathbf{C}}^{\pm}$ are not properly matched, and the boundary conditions are not necessarily satisfiable.

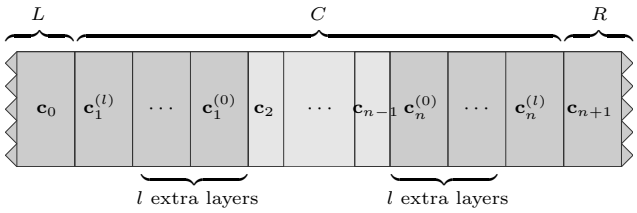


Figure 5.3: Two-probe system in which the C region boundaries are expanded by l extra layers of the corresponding connecting electrode.

5.4 Inserting extra electrode layers

In order to diminish the errors introduced by excluding evanescent modes, we propose to insert additional electrode layers in the central region as illustrated schematically in Fig. 5.3. As discussed above, this would quickly reduce the imprint of the rapidly decaying evanescent modes in the boundary layer wave functions $\mathbf{c}_1$ and $\mathbf{c}_n$, which means that the critical components outside the column spaces $\mathbf{C}^{\pm}$ become negligible at an exponential rate in terms of the number of additional layers. We emphasize that the inserted layers may be "fictitious" in the sense that they can be accommodated by block Gaussian elimination operations prior to solving Eq. (3.91) for the original system.

Let us now analyze the above statements formally. In the particular case, where l extra principal layers of the connecting electrodes are inserted and also the border layers of the C region are identical to the connecting electrode layers, we can write the boundary matching equations as

$$\mathbf{c}_0 = (\tilde{\mathbf{B}}_L^+)^{-1}\mathbf{c}_1^{(l)+} + (\tilde{\mathbf{B}}_L^-)^{-1}\mathbf{c}_1^{(l)-} \tag{5.6}$$

for the left boundary and

$$\mathbf{c}_{n+1} = \tilde{\mathbf{B}}_R^+\mathbf{c}_n^{(l)+} + \tilde{\mathbf{B}}_R^-\mathbf{c}_n^{(l)-} \tag{5.7}$$

for the right boundary, where $\mathbf{c}_1^{(l)+} = \lambda_{L,k}^+\mathbf{c}_{L,k}^+$ and $\mathbf{c}_n^{(l)-} = \mathbf{0}$ are fixed as boundary conditions. We point out, that the l extra layers are bulk layers extending from each electrode and therefore connected via the relation in Eq. (5.5) for L and R, respectively. Moreover, since the electrode wave functions can always be expanded in the corresponding complete set of bulk modes, we may write

$$\mathbf{c}_i^{\pm} = \mathbf{C}^{\pm}\mathbf{a}_i^{\pm} = [\tilde{\mathbf{C}}^{\pm}, \mathring{\mathbf{C}}^{\pm}] \begin{bmatrix} \tilde{\mathbf{a}}_i^{\pm} \\ \mathring{\mathbf{a}}_i^{\pm} \end{bmatrix}, \tag{5.8}$$

where $\mathbf{a}_i^{\pm} = [\tilde{\mathbf{a}}_i^{\pm T}, \mathring{\mathbf{a}}_i^{\pm T}]^T$ are vectors that contain the expansion coefficients and subscript L is implied for the left electrode ($i \leq 1$), and R for the right electrode

($i \geq n$). Thus we may consider the (unfixed) boundary wave functions entering Eqs. (5.6) and (5.7), by explicitly writing

$$\mathbf{c}_1^{(l)-} = (\mathbf{B}_L^-)^{-l}\mathbf{c}_1^- = [\tilde{\mathbf{C}}_L^-, \mathring{\mathbf{C}}_L^-] \begin{bmatrix} (\tilde{\mathbf{\Lambda}}_L^-)^{-l}\tilde{\mathbf{a}}_1^- \\ (\mathring{\mathbf{\Lambda}}_L^-)^{-l}\mathring{\mathbf{a}}_1^- \end{bmatrix}, \tag{5.9}$$

and

$$\mathbf{c}_n^{(l)+} = (\mathbf{B}_R^{\pm})^l\mathbf{c}_n^+ = [\tilde{\mathbf{C}}_R^+, \mathring{\mathbf{C}}_R^+] \begin{bmatrix} (\tilde{\mathbf{\Lambda}}_R^+)^{l}\tilde{\mathbf{a}}_n^+ \\ (\mathring{\mathbf{\Lambda}}_R^+)^{l}\mathring{\mathbf{a}}_n^+ \end{bmatrix}, \tag{5.10}$$

using the definition $\mathbf{B}^{\pm} = \mathbf{C}^{\pm}\mathbf{\Lambda}^{\pm}(\mathbf{C}^{\pm})^{-1}$. This shows that the critical components outside the column spaces of $\tilde{\mathbf{C}}_L^{\pm}$ and $\tilde{\mathbf{C}}_R^{\pm}$ are given by coefficients $(\mathring{\mathbf{\Lambda}}_L^-)^{-l}\mathring{\mathbf{a}}_1^-$ and $(\mathring{\mathbf{\Lambda}}_R^+)^{l}\mathring{\mathbf{a}}_n^+$, respectively. Assuming that we exclude fastest decaying of the evanescent modes according to Eq. (5.1), we will have $|\lambda_k| > \lambda_{\min}^{-1}$ for the diagonal elements of $\mathring{\mathbf{\Lambda}}_L^-$ and $|\lambda_k| < \lambda_{\min}$ for the diagonal elements of $\mathring{\mathbf{\Lambda}}_R^+$. More importantly, since $\lambda_{\min}$ is always less than 1, these coefficients always decrease as a function of l.

We therefore conclude that WFM with the reduced Bloch matrices $\tilde{\mathbf{B}}^{\pm}$ approaches the exact case with $\mathbf{B}^{\pm}$ if additional electrode layers are inserted as suggested, and thus, that the solution $\mathbf{c}_C$ obtained from Eq. (3.91) when only a reduced set of bulk modes are used, approaches the correct solution accordingly.

5.5 Accuracy and error analysis

As pointed out above, the exclusion of some of the evanescent modes from the mode matrices $\mathbf{C}^{\pm}$ may introduce errors because the column spaces in $\tilde{\mathbf{C}}^{\pm}$ are incomplete. However, it is not obvious to which extend this influences the accuracy of the transmission and reflection calculations from the scattering states solutions $\mathbf{c}_1^{(l)-}$ and $\mathbf{c}_n^{(l)+}$. It is therefore important to be able to estimate and monitor the accuracy of the results obtained. We now discuss how this can be done in a systematic fashion in terms of the parameter $\lambda_{\min}$ and the number l of extra electrode layers.

Consider first the accuracy of the transmission matrix $\mathbf{t}$ in the case of the extended two-probe system in Fig. 5.3. Initially, for a specific incoming mode k, we would like to compare the correct result obtained with the complete set of modes (cf. Eq. (3.83)),

$$\mathbf{t}_k = \begin{bmatrix} \tilde{\mathbf{t}}_k \\ \mathring{\mathbf{t}}_k \end{bmatrix} = [\tilde{\mathbf{C}}_R^+, \mathring{\mathbf{C}}_R^+]^{-1}\mathbf{c}_n^{(l)+}, \tag{5.11}$$

to the result obtained with the reduced mode matrix (denoted by a prime),

$$\mathbf{t}'_k = \begin{bmatrix} \tilde{\mathbf{t}}'_k \\ \mathring{\mathbf{0}}' \end{bmatrix} = [\tilde{\mathbf{C}}_R^+, \mathring{\mathbf{0}}]^{-1}\mathbf{c}_n^{(l)+}, \tag{5.12}$$

where $\mathring{\mathbf{0}}'$ and $\mathring{\mathbf{0}}$ represents the zero vector and zero matrix of size $\mathring{m}_R$ and $m_R \times \mathring{m}_R$, respectively.

Notice that the important coefficients in $\mathbf{t}_k$ and $\mathbf{t}'_k$ for transmission calculations are the ones representing the Bloch modes which enters the Landauer-Büttiker formula in Eq. (3.8). Since these are never excluded they will always be located within the first $\tilde{m}_R$ elements, i.e., in $\tilde{\mathbf{t}}_k$ and $\tilde{\mathbf{t}}'_k$. It then suffices to compare these parts of the transmission matrix which we can do as follows.

From the properties of the pseudo inverse we are able to write the relation

$$(\tilde{\mathbf{C}}_R^+)^{-1}[\tilde{\mathbf{C}}_R^+, \mathring{\mathbf{C}}_R^+] = [\tilde{\mathbf{I}}, (\tilde{\mathbf{C}}_R^+)^{-1}\mathring{\mathbf{C}}_R^+], \tag{5.13}$$

where $\tilde{\mathbf{I}}$ is the identity matrix of order equal to the number of included modes $\tilde{m}_R$. Using the expression in Eq. (5.10) it then follows that

$$\tilde{\mathbf{t}}_k = (\tilde{\mathbf{\Lambda}}_R^+)^l \tilde{\mathbf{a}}_n^+, \tag{5.14}$$

and

$$\tilde{\mathbf{t}}'_k = \tilde{\mathbf{t}}_k + (\tilde{\mathbf{C}}_R^+)^{-1}\mathring{\mathbf{C}}_R^+(\mathring{\mathbf{\Lambda}}_R^+)^l \mathring{\mathbf{a}}_n^+, \tag{5.15}$$

where the $\tilde{\mathbf{t}}'_k$ expression clearly corresponds to the correct coefficients $\tilde{\mathbf{t}}_k$ plus an error term.

We have already established in the previous section that the $(\mathring{\mathbf{\Lambda}}_R^+)^l \mathring{\mathbf{a}}_n^+$ factor in the error term will decrease as a function of l. To ascertain that the total error term also decreases, we look at the 2-norm of $(\tilde{\mathbf{C}}_R^+)^{-1}\mathring{\mathbf{C}}_R^+$, which satisfies

$$||(\tilde{\mathbf{C}}_R^+)^{-1}\mathring{\mathbf{C}}_R^+||_2 \leq \mathring{m}_R^{\frac{1}{2}} ||(\tilde{\mathbf{C}}_R^+)^{-1}||_2, \tag{5.16}$$

since $||\mathring{\mathbf{C}}_R^+||_2 \leq \mathring{m}_R^{\frac{1}{2}}$ when all evanescent modes are assumed to be normalized (see Eq. (3.87)). The norm $||(\tilde{\mathbf{C}}_R^+)^{-1}||_2$ can be readily evaluated and depends on the set of modes included via the parameter $\lambda_{\min}$ but not on l. We then have that $(\tilde{\mathbf{C}}_R^+)^{-1}\mathring{\mathbf{C}}_R^+$ is independent of l, and consequently, that the error term in Eq. (5.15) must decrease as a function of l.

5.5.1 Error estimates

We now derive expressions in order to estimate and monitor the error in a calculation of the total transmission coefficient $T(E)$ with the proposed WFM method. Writing Eq. (5.15) as $\tilde{\mathbf{t}}'_k = \tilde{\mathbf{t}}_k + \tilde{\boldsymbol{\epsilon}}_k$, where $\tilde{\boldsymbol{\epsilon}}_k$ holds the errors on the coefficients of the kth column, we can insert this in the Landauer-Büttiker formula in Eq. (3.7) and obtain an expression for $T'(E)$, given by

$$T'(E) = T(E) + \sum_{kk'} (\tilde{t}^*_{kk'}\tilde{\epsilon}_{kk'} + \tilde{\epsilon}^*_{kk'}\tilde{t}_{kk'} + |\tilde{\epsilon}_{kk'}|^2) \tag{5.17}$$

where $T(E)$ is the exact result and the summation is over the Bloch modes k and k' in the left and right electrode, respectively.

In the attempt to estimate the order of the error term in Eq. (5.17), we may (as a worst case approximation) take all diagonal elements of $\mathring{\mathbf{\Lambda}}_R^+$ to be equal to the maximum range $\lambda_{\min}$ of Eq. (5.1), which makes all elements $\tilde{\epsilon}_{kk'}$ proportional to $\lambda_{\min}^l$. Thus we arrive at the simple relation

$$|T' - T| \sim \lambda_{\min}^l + O((\lambda_{\min}^l)^2), \tag{5.18}$$

which shows that the error decreases exponentially in terms of the number of extra layers l. We will adopt the expression in Eq. (5.18) as a reasonable order of magnitude estimate of the accuracy of $T'(E)$.

Alternatively, if we assume that the magnitude of the error on the right-going and left-going components are of similar order, we can attempt to monitor the error arising on the boundary conditions. In order to do this, we introduce the new coefficient vectors

$$\tilde{\mathbf{b}}_{L,k} = (\tilde{\mathbf{C}}_R^+)^{-1}(\mathbf{c}_1^{(l)+} - \lambda_{L,k}^+ \mathbf{c}_{L,k}^+) \tag{5.19}$$

and

$$\tilde{\mathbf{b}}_{R,k} = (\tilde{\mathbf{C}}_R^-)^{-1}\mathbf{c}_n^{(l)-}. \tag{5.20}$$

We note that $|\tilde{\mathbf{b}}_{L,k}| = 0$ and $|\tilde{\mathbf{b}}_{R,k}| = 0$ in the case where the boundary conditions are exactly satisfied (i.e., $\mathbf{c}_1^{(l)+} = \lambda_{L,k}^+ \mathbf{c}_{L,k}^+$ and $\mathbf{c}_n^{(l)-} = \mathbf{0}$). Thus $\tilde{\mathbf{b}}_{R,k}$, for example, represents the error on the left-going components within the right boundary layer in the same way that $\tilde{\boldsymbol{\epsilon}}_k$ represents the error on the right-going (transmitted) components. We would therefore expect the same order of magnitude of $|\tilde{\mathbf{b}}_{R,k}|$ and $|\tilde{\boldsymbol{\epsilon}}_k|$ in an actual calculation for a given mode k.

This suggests another order of magnitude accuracy estimate for $T(E)$, which is straightforward to monitor using the results available with the reduced set of bulk modes. By relating $|\tilde{\mathbf{b}}_{R,k}| \sim |\tilde{\boldsymbol{\epsilon}}_k|$ and using Eq. (5.17), we can write

$$|T' - T| \leq \sum_k (2|\tilde{\mathbf{t}}_k||\tilde{\boldsymbol{\epsilon}}_k| + |\tilde{\boldsymbol{\epsilon}}_k|^2) \sim \sum_k (2|\tilde{\mathbf{t}}_k||\tilde{\mathbf{b}}_{R,k}| + |\tilde{\mathbf{b}}_{R,k}|^2), \tag{5.21}$$

where all the vector norms (e.g., $|\tilde{\mathbf{t}}_k|^2 = \sum_{k'} |\tilde{t}_{kk'}|^2$) are assumed to be taken over the elements corresponding to propagating bulk modes k' only.

We note without explicit derivation that similar arguments for the reflection matrix with columns $\tilde{\boldsymbol{r}}'_k$ and the total reflection coefficient R', as presented above for $\tilde{\mathbf{t}}'_k$ and T', results in the same accuracy expressions for $|R' - R|$ as for $|T' - T|$ in Eqs. (5.18) and (5.21), if we substitute $\tilde{\mathbf{t}}_k \to \tilde{\mathbf{r}}_k$ and $\tilde{\mathbf{b}}_{R,k} \to \tilde{\mathbf{b}}_{L,k}$.

5.6 Implementation

We now turn to the practical implementation of the new method. To begin with we note that in order to benefit the most from the approach of excluding

evanescent modes, we have to be able to target and compute *only* the modes of interest. This is in contrast to the standard method of evaluating all modes and subsequently throwing away a (relatively modest) subset of them (see Sec. 3.3.2).

The resulting algorithm which we have implemented is written,

ALGORITHM X: *Efficient WFM method*

1. obtain $\tilde{\mathbf{C}}_L^{\pm}, \tilde{\mathbf{\Lambda}}_L^{\pm}$ from QEP (see Chap. 6)
2. $[\tilde{\mathbf{Q}}_L^{\pm}, \tilde{\mathbf{R}}_L^{\pm}] = \mathrm{QR}\{\tilde{\mathbf{C}}_L^{\pm}\}$, solve $\tilde{\mathbf{R}}_L^{\pm}\tilde{\mathbf{C}}_L^{\pm} = (\tilde{\mathbf{Q}}_L^{\pm})^{\dagger}$
3. $\tilde{\mathbf{B}}_L^{\pm} := \tilde{\mathbf{C}}_L^{\pm}\tilde{\mathbf{\Lambda}}_L^{\pm}\tilde{\mathbf{C}}_L^{\pm}$
4. $\tilde{\mathbf{\Sigma}}_L := \bar{\mathbf{H}}_{L,L}^{\dagger}[\bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}^{\dagger}(\tilde{\mathbf{B}}_L^{-})^{-1}]^{-1}\bar{\mathbf{H}}_{L,L}$

} the left electrode

5. obtain $\tilde{\mathbf{C}}_R^{+}, \tilde{\mathbf{\Lambda}}_R^{+}$ from QEP (see Chap. 6)
6. $[\tilde{\mathbf{Q}}_R^{+}, \tilde{\mathbf{R}}_R^{+}] = \mathrm{QR}\{\tilde{\mathbf{C}}_R^{+}\}$, solve $\tilde{\mathbf{R}}_R^{+}\tilde{\mathbf{C}}_R^{+} = (\tilde{\mathbf{Q}}_R^{+})^{\dagger}$
7. $\tilde{\mathbf{B}}_R^{+} := \tilde{\mathbf{C}}_R^{+}\tilde{\mathbf{\Lambda}}_R^{+}\tilde{\mathbf{C}}_R^{+}$
8. $\tilde{\mathbf{\Sigma}}_R := \bar{\mathbf{H}}_{R,R}[\bar{\mathbf{H}}_R + \bar{\mathbf{H}}_{R,R}\tilde{\mathbf{B}}_R^{+}]^{-1}\bar{\mathbf{H}}_{R,R}^{\dagger}$

} the right electrode

9. $\tilde{\mathbf{q}}_1' = -[\bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{C}_L^{+} + \tilde{\mathbf{\Sigma}}_L\tilde{\mathbf{C}}_L^{+}\tilde{\mathbf{\Lambda}}_L^{+}]\mathbf{a}^{\mathrm{in}}$
10. initialize $\mathbf{A}_1' := \bar{\mathbf{H}}_{11} - \mathbf{\Sigma}_L,\ \mathbf{A}_{\mathrm{n}}^{\backslash} := \bar{\mathbf{H}}_{\mathrm{n,n}}$
11. **for** $i := 1, \ldots, l$
12. solve $\bar{\mathbf{H}}_{L,L} = \mathbf{X}\mathbf{A}_1'$ for $\mathbf{X}$
13. $\mathbf{A}_1' := \bar{\mathbf{H}}_L - \mathbf{X}\bar{\mathbf{H}}_{L,L}$
14. $\mathbf{q}_1' := -\mathbf{X}\mathbf{q}_1'$
15. **end**

} extra layers left

16. execute downwards sweep of ALGORITHM V for i = 1, . . . , n

17. **for** $i := 1, \ldots, l$
18. solve $\bar{\mathbf{H}}_{L,L} = \mathbf{X}\mathbf{A}_{\mathrm{n}}'$ for $\mathbf{X}$
19. $\mathbf{A}_{\mathrm{n}}' := \bar{\mathbf{H}}_L - \mathbf{X}\bar{\mathbf{H}}_{L,L}$
20. $\mathbf{q}_{\mathrm{n}}' := -\mathbf{X}\mathbf{q}_{\mathrm{n}}'$
21. **end**

} extra layers right

22. $\mathbf{A}_{\mathrm{n}}^{\vee} := \mathbf{A}_{\mathrm{n}}' - \mathbf{\Sigma}_R$
23. repeat lines $11 - 21$ in "reverse" (i.e., extended upwards sweep)

24. **for** $i := 1, \ldots, \mathrm{n}$
25. solve $\bar{\mathbf{H}}_i'\mathbf{c}_i = \mathbf{q}_i^{\vee}$
26. **end**

} scattering states solution

(5.22)

Notice that the extra layers have been inserted "fictitiously" in the sense that they are implemented by block Gaussian elimination operations before and after

the standard downwards and upwards sweeps.

Since the majority of the evanescent modes can be excluded from the WFM calculation altogether by using ALGORITHM X, the numerical operations in lines 1-9 are in general much less costly in comparison with lines 1-9 of the original WFM method in ALGORITHM V. For many large two-probe systems, which the current method is aimed for, this algorithm can therefore be very efficient in combination with an appropriate iterative eigenvalue problem solver in steps 1 and 5. A particular procedure based on the restarted Arnoldi method developed in the next chapter is adopted in the following applications.

5.7 Examples

We exemplify the previous discussion quantitatively by looking at two cases. First, in order demonstrate of applicability of the derived error estimates in Eqs. (5.18) and (5.21), we consider the Au–CNT(8,0)×4–Au system. Second, we return to the motivating arguments in Sec. 5.1 and compute the speed-ups achieved by the proposed WFM method for the systems mentioned there.

5.7.1 Benchmarking example

Consider the Au–CNT(8,0)×4–Au two-probe system in the right part of Fig. 5.1, consisting of the Au(111) electrode described earlier, and a 128 atom (4 unit cells) device of zigzag-(8,0) carbon nanotube (CNT). For energy $E = -1.0$ eV and $E = -1.5$ eV, we have calculated the deviation between the total transmission obtained when all bulk modes are taken into account (T) and when evanescent modes are excluded (T') as specified with different settings of $\lambda_{\min}$. Deviations are also determined for the corresponding total reflection coefficients (R and R'). Fig. 5.4 shows the results as a function of l, together with the estimate $\lambda^l_{\min}$ of Eq. (5.18) and the estimate of Eq. (5.21) both for the transmission and reflection coefficients, where the higher order terms have been neglected,

We observe that the absolute error in the obtained transmission coefficients (red curves) and reflection coefficients (blue curves) are generally decreasing as a function of l, following the same or better convergence rate as $\lambda^l_{\min}$ (dashed line). Looking closer at results for neighbor l values, we see that the errors in the $E = -1.5$ eV case initially exhibit wave-like oscillations. This is directly related to the wave form of the evanescent modes that have been excluded (see the propagation of the slowest decaying black curves in Fig. 5.2(b)), since the representation of these modes in the reduced spaces $\tilde{\mathbf{C}}^{\pm}$ (i.e., the expansion coefficients in $\tilde{\boldsymbol{\epsilon}}_k$) may shift when l is increased. In other words, although the norm of the errors $|\tilde{\boldsymbol{\epsilon}}_k|$ are decreasing as a function of l, the specific error $\tilde{\epsilon}_{kk'}$ on a given (large) coefficient of $\tilde{t}'_{kk'}$ or $\tilde{r}'_{kk'}$ may increase, which means that the overall error term in Eq. (5.17) can go up. Fortunately, however, this is only a local phenomenon with the global trend being rapidly decreasing errors.

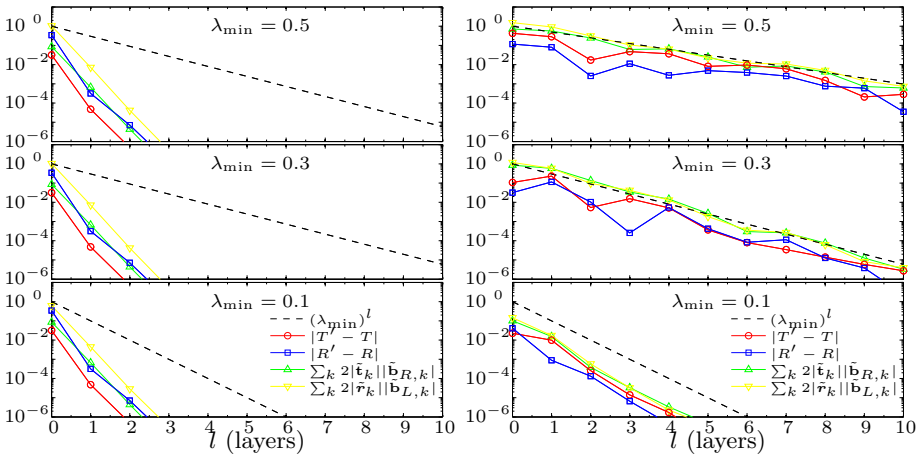


Figure 5.4: Error (absolute) in the calculated total transmission and reflection coefficients T' and R' as a function of l. The panels show the cases of $\lambda_{\min}$ set to 0.5, 0.3 and 0.1, for energy $E = -1.0$ eV (left) and $E = -1.5$ eV (right) The dashed line indicate the theoretical accuracy estimate $\lambda_{\min}^l$.

It is striking that all the results for $E = -1.0$ eV in the left part of Fig. 5.4 are completely identical for all settings of $\lambda_{\min}$. This indicates that there are no modes located in the area $0.1 \leq |\lambda_k| \leq 0.5$ of the unit disc at this energy. Moreover, the quality of the simple accuracy estimate $\lambda_{\min}^l$ and the estimates expressed by Eq. (5.21) for the transmission coefficients (green curves) and reflection coefficients (yellow curves), respectively, are very distinct. For relatively large $\lambda_{\min}$ in the $E = -1.5$ eV case, all estimates are very good. However, for $E = -1.0$ eV or for smaller values of $\lambda_{\min}$, only the latter two retain a high quality while the $\lambda_{\min}^l$ estimate tends to be overly pessimistic. It is important to remember, that these estimates are by no means strict conditions but very reasonable to make an order of magnitude estimate of the accuracy.

We note in passing, that the results in the top panels of Fig. 5.4 corresponds to using *only* the propagating Bloch modes in the transmission calculation (i.e., only 3 out of 243 for $E = -1.5$ eV, cf. Fig. 5.2). Still we are able to compute T and R to an absolute accuracy of three digits by inserting 2×5 extra electrode layers in the two-probe system. This is quite remarkable and shows promise for large-scale systems, e.g., with nano-wire electrodes, for which the total number of evanescent modes available becomes exceedingly great.

Table 5.2: CPU-times in seconds and speed-up when using the efficient WFM method in ALGORITHM X for calculating $\boldsymbol{t}$ and $T(E)$ at 20 different energies inside $E \in [-2\text{ eV}; 2\text{ eV}]$ for various two-probe systems (cf. Table 5.1). Columns 4-5 show the percentage of the CPU-time used for computing the electrode bulk modes from the QEP vs. solving the central region linear systems in Eq. (3.91).

System	Atoms	CPU	Speed-up	QEP	Eq. (3.91)
Li–Li	32(8)	0.1	1.1	71%	28%
Fe–MgO–Fe	27(6)	2.1	1.0	55%	44%
Al–C×7–Al	74(18)	2.3	1.7	71%	27%
Au–DTB–Au	102(27)	42.0	2.3	67%	32%
Au–CNT(8,0)×1–Au	140(27)	34.0	2.6	50%	49%
Au–CNT(8,0)×5–Au	268(27)	68.4	1.9	26%	73%
CNT(8,0)–CNT(8,0)	192(64)	31.5	4.3	61%	37%
CNT(4,4)–CNT(8,0)	256(64|64)	33.6	3.8	62%	37%
CNT(5,0)–CNT(10,0)	300(40|80)	54.4	2.5	41%	58%
CNT(18,0)–CNT(18,0)	576(144)	469.7	3.3	48%	51%

5.7.2 Speed-up

In conclusion to this section it is appropriate to demonstrate that the proposed WFM method, which we have implemented as ALGORITHM X, is indeed both robust and efficient. We have therefore calculated the transmission spectrum $T(E)$ for 20 different energies inside $E \in [-2\text{ eV}; 2\text{ eV}]$ for a selection of frequently studied two-probe systems using parameters $\lambda_{\min} = 0.1$ and $l = 1$. The CPU-times for the efficient WFM method is given in Table 5.2 and can be directly compared with the CPU-times for the standard WFM method listed in Table 5.1. The actual transmission spectra produced by applying the two methods are identical to more than three significant digits (for explicit comparison of $T(E)$ curves, see Sec. 6.4). Also, we have confirmed the results in the majority of the cases[1] by reviewing the original publications that consider these systems [67, 40, 41, 66, 68]. We choose therefore not to show the $T(E)$ spectra obtained.

The speed-ups achieved by the efficient WFM method for the different two-probe systems are displayed in the fourth column of Table 5.2. They range from 1 to 4.3 and follow a common trend; the larger the electrode unit cells are, both in general and relative to the central region, the better speed-up can be expected. Also the percentages of time spent in the key stages can be directly compared to the corresponding numbers for the standard method (cf. Table 5.1). It is apparent that the relative cost has shifted from the electrode QEPs to the solution of the central region linear system in Eq. (3.91). This is partly due to

[1] Fe–MgO–Fe, Al–C×7–Al, Au–DTB–Au, Au–CNT(8,0)×N–Au, CNT(8,0)–CNT(8,0).

the extra ("fictitious") layers in the C region, and partly due to the significantly cheaper evaluation of the self-energy matrices (see next chapter).

We refer the reader to **PAPER II** and Sec. 6.4 for more verification that the above WFM method calculates the correct scattering states and transmission spectrum in an efficient manner.

Chapter 6

Krylov subspace method for computing self-energy matrices

In the previous chapters of this thesis, we have given detailed descriptions of the Green's function method and the WFM method for the modeling quantum transport in two-probe nano-scale devices. We have seen that in both methods, it is necessary to evaluate the self-energy matrices of each electrode for a number of different energies. For most systems this represents the dominant part of the computational work. So far, the fastest method for obtaining the self-energy matrices has been via the solutions of the bulk QEPs of the electrodes. As illustrated in the previous chapter, only the propagating and the slowly decaying evanescent states in the bulk electrodes contribute to the transmission of electrons through a two-probe device of some extension. These states correspond to the solutions of the QEP that have complex eigenvalues in the vicinity of the unit circle. One can then generate reduced self-energy matrices on the basis of a few selected solutions of the QEP, which include all the electrode-device coupling information of interest. To exploit this in practice, an algorithm to search for and compute the desired quadratic eigenpairs is required. In this chapter, we develop such a method using an iterative Krylov subspace technique.

6.1 Introduction

With the recent surge of very efficient iterative schemes to obtain a few specific eigenpairs of large-scale eigenvalue problems (see Refs. [69, 70, 71, 72, 73] and references therein), we believe such techniques to be the natural starting point for solving the electrode QEPs in Eq. (3.56). In particular the methods that are tailored for quadratic and polynomial matrix problems seem to bee promising [74, 75, 76, 77]. In such methods, the explicit linearization of the QEP as performed in Eq. (3.58) and the subsequent doubling of the problem size, can

be effectively avoided [77]. Furthermore, by applying appropriate restarting mechanisms (see, e.g., [69]) as part of the iterative schemes, one is able to determine the selected eigenpairs to a given accuracy at a relatively low cost.

The most difficult aspect of applying the existing iterative methods for the QEPs at hand, is the particular interior location of the desired eigenvalues within the complex eigenvalue spectrum. In many of the applications for which large-scale iterative methods are used it is the extreme (i.e., exterior) of the spectrum that is of interest. In our case, we require the specific eigenpairs with eigenvalues close to the unit circle. Fortunately, the Arnoldi method combined with a shift-and-invert strategy has proven to be an effective tool in obtaining selected interior eigenvalues of large-scale general complex eigenproblems.[70, 71, 72]. Other iterative methods, which are also suitable for this task, is the folded spectrum variant of the preconditioned conjugate gradient method [78] and the Jacobi-Davidson method for QEPs [79, 80, 73]. We here adopt the Arnoldi method, which is the simplest and most straightforward to implement, and leave the study of the applicability of the latter two methods to future work.

6.1.1 Arnoldi procedure

The Arnoldi method was first introduced as a direct algorithm for reducing a general matrix to upper Hessenberg form [81]. It was later discovered to be an excellent iterative technique for finding eigenpairs of large matrices [82]. The procedure can be essentially viewed as a modified Gram-Schmidt process for constructing an orthogonal basis of a Krylov subspace. It works for general non-Hermitian matrices as is required in our case. We begin here with a brief presentation of the basic procedure.

The Krylov subspace of dimension m generated by an $n \times n$ matrix $\mathbf{A}$ and an initial vector $\mathbf{v}_1$, is given by

$$\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1) \equiv \text{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \mathbf{A}^2\mathbf{v}_1, \ldots, \mathbf{A}^{m-1}\mathbf{v}_1\}, \tag{6.1}$$

which is the span of the vectors available from the power method [83]. In order to determine the Krylov subspace, we apply the Arnoldi procedure which generates an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ for $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$. ALGORITHM XI below lists the steps of a continuable version of the Arnoldi procedure which is initially

called with a parameter $k = 1$ and a random starting vector $\mathbf{v}_1$.

ALGORITHM XI: *Arnoldi procedure (continuable)*

1. **if** $k = 1,\ \mathbf{v}_1 = \mathbf{v}_1/||\mathbf{v}_1||_2$
2. **for** $j = k, k+1, \ldots, m$ **do**
3. $\quad \mathbf{v} = \mathbf{A}\mathbf{v}_j$
4. $\quad$ **for** $i = 1, 2, \ldots, j$**do**
5. $\quad\quad \mathrm{H}_{ij} = \mathbf{v}_i^T\mathbf{v}$
6. $\quad\quad \mathbf{v} = \mathbf{v} - \mathrm{H}_{ij}\mathbf{v}_i$
7. $\quad$ **end**
8. $\quad \mathrm{H}_{j+1,j} = ||\mathbf{v}||_2$
9. $\quad$ **if** $\mathrm{H}_{j+1,j} = 0, m = j,$ **breakdown**
10. $\quad \mathbf{v}_{j+1} = \mathbf{v}/h_{j+1,j}$
11. **end**

(6.2)

After $m-1$ iterations an $n \times m$ matrix whose columns are the orthonormal basis vectors for $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$ is available,

$$\mathbf{V}_m = \big(\mathbf{v}_1, \ldots, \mathbf{v}_m\big). \tag{6.3}$$

The projection of the matrix $\mathbf{A}$ onto $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$ is then

$$\mathbf{H}_m = \mathbf{V}_m^\dagger \mathbf{A}\mathbf{V}_m, \tag{6.4}$$

where $\mathbf{H}_m$ is $m \times m$ and upper Hessenberg (i.e., it has zeros below its lower bidiagonal). The matrix $\mathbf{H}_m$ (not to be confused with Hamiltonian matrix $\mathbf{H}_i$) is also constructed by ALGORITHM XI. Approximate solutions of the eigenproblem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ can subsequently be obtained as the so-called Ritz eigenpairs $(\gamma, \mathbf{V}_m\mathbf{y})$ of the projected eigenproblem $\mathbf{H}_m\mathbf{y} = \gamma\mathbf{y}$. More specifically, as m increases the Ritz pairs becomes increasingly better approximations to certain eigenpairs of $\mathbf{H}_m$ (we point to Refs. [83, 80] for details).

We note that ALGORITHM XI stops prematurely at line 9 if the length of the current vector to be added to $\mathbf{V}$ is zero. If this is the case, then the projection of $\mathbf{A}$ onto the current subspace of dimension j will be exact. Therefore such a breakdown is called a "lucky breakdown". In any case, the algorithm produces the output matrices $\mathbf{V}_{m+1}$ and $\mathbf{H}_{m+1}$ which by construction will satisfy the relation

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_{m+1}\underline{\mathbf{H}}_{m+1}, \tag{6.5}$$

where $\mathbf{V}_{m+1} = (\mathbf{V}_m, \mathbf{v}_{m+1})$ is a rectangular $n \times (m+1)$ matrix and $\underline{\mathbf{H}}_{m+1} = (\mathbf{H}_m^T, (0, \ldots, 0, h_{m+1,m})^T)^T$ is an $(m+1) \times m$ upper Hessenberg matrix.

6.2 Krylov subspace algorithm

In this section we describe the new Krylov subspace method for evaluating the electrode self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$. The crucial assumption in the approach is that we may strip the less important modes from the mode matrices $\mathbf{C}^{\pm}$ in Eq. (3.62) and still obtain a good approximation to the self-energy matrix as described in Chap. 5. Our current method, which targets the specific modes that are most important, can be characterized as a shift-and-invert Arnoldi method with adaptive subspace size. We will describe the key ingredients of the method in the following. The goal is to present an alternative for obtaining the self-energy matrices, which is faster than existing techniques.

6.2.1 Shift-and-invert transformations

Iterative methods based on Krylov subspaces produce Ritz values that converge fastest to the dominant part of the eigenvalue spectrum given by the extremal eigenvalues [83]. In the current application, it is the interior of the eigenvalue spectrum that is of interest, in particular the eigenvalues λ_k that satisfy $\lambda_{\min} \leq |\lambda_k| \leq \lambda_{\min}^{-1}$. To be able to find this part of the spectrum efficiently, we employ a shift-and-invert strategy which implies that the QEP in Eq. (3.56) is rewritten as

$$\big(\mu^2\mathbf{M} + \mu\mathbf{C} + \mathbf{K}\big)\mathbf{c}_k = 0, \tag{6.6}$$

where

$$\mathbf{M} = \bar{\mathbf{H}}_{L,L}^{\dagger} + \sigma\bar{\mathbf{H}}_L + \sigma^2\bar{\mathbf{H}}_{L,L}, \tag{6.7}$$

$$\mathbf{C} = \bar{\mathbf{H}}_L + 2\sigma\bar{\mathbf{H}}_{L,L}, \tag{6.8}$$

$$\mathbf{K} = \bar{\mathbf{H}}_{L,L}, \tag{6.9}$$

and

$$\mu = \frac{1}{\lambda_k - \sigma}. \tag{6.10}$$

Notice that we have left out the implied subscript L on the eigenvalues $\lambda_{L,k}$ and eigenvectors $\mathbf{c}_{L,k}$ (we do this from here on). By transforming the QEP, the eigenvalues of Eq. (3.56) have been shifted by σ and inverted while the eigenvectors are unchanged. Thus the dominant part of the spectrum of Eq. (6.6) now corresponds to the eigenvalues of the original QEP closest to the shift σ.

As discussed in Sec. 3.3.2, the simplest technique for solving the QEP in Eq. (6.6) is by linearizing it to a generalized eigenvalue problem of twice the size, see Eq. (3.58). However, for the values of σ used in this work, $\mathbf{M}$ is always well-conditioned. A linearization is therefore possible as a standard eigenvalue problem of size $2m_L$, given in the current notation as (cf. Eq. (3.57))

$$\mathbf{A}\mathbf{x} = \mu\mathbf{x}, \tag{6.11}$$

where $\mathbf{A}$ is

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \boldsymbol{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C} \end{pmatrix}, \tag{6.12}$$

and the $2m_L$ eigenvalues μ_k are identical to the ones of Eq. (6.6). The eigenvectors of Eq. (3.58) are given by $\mathbf{x}_k^T = (\mathbf{c}_k, \mu_k \mathbf{c}_k^T)$, so that the original eigenvectors $\mathbf{c}_k$ can be selected as the first m_L elements of $\mathbf{x}_k$.

If we assume that the Hamiltonian and overlap matrices for the electrodes are real, then the spectrum of the QEP in Eq. (3.56) is symmetric with respect to the real axis of the complex plane and the eigenvalues are either real or occur in complex conjugate pairs [63]. In addition, as seen by transposing Eq. (3.56), the eigenvalues in this case also come in pairs λ_k and $1/\lambda_k$. We will use these properties to present a simplified method for the extraordinary case of real $\bar{\mathbf{H}}_L$ and $\bar{\mathbf{H}}_{L,L}$, and subsequently discuss the steps required for the general complex case in Sec. 6.2.5. The purpose of the current method is thus to determine the eigenpairs $(\lambda_k, \mathbf{c}_k)$ of Eq. (3.56) that satisfy $\lambda_{\min} \leq |\lambda_k| \leq 1$, for $\lambda_{\min} > 0$ (the pairs that satisfy $1 \leq |\lambda_k| \leq \lambda_{\min}^{-1}$ can subsequently be obtained as $(\lambda_k^{-1}, \mathbf{c}_k)$).

As is apparent from the polar plot example in Fig. 5.2, the majority of the eigenvalues with $|\lambda_k| \leq 1$ are located near the origin. Therefore, it is not efficient to apply the shift $\sigma = 0$ in order to obtain the wanted eigenvalues which lie in the outskirts of the unit disc. Instead we may apply four different shifts, given by $\sigma = \pm 1/\sqrt{2}$ and $\sigma = \pm \hat{\imath}/\sqrt{2}$, in four separate Arnoldi procedures. Each of these then cover a quarter-slice of the unit disc and produce Ritz values that converge fast to eigenvalues close to the given shift. Simple sorting techniques can be employed in each Arnoldi procedure to take only the portion of the Ritz pairs into account that is covered by a given shift.

When applying the shift-and-invert strategy devised, two of the shifts have to be complex. In practice this means working in complex arithmetic or doubling the size of the problem [84]. However, in the case of real Hamiltonians it is advantageous to search for the complex eigenvalues in conjugate pairs and thereby eliminate one of the complex shifts. Moreover, this can be done almost entirely in real arithmetic as follows.

Notice that Eq. (6.11) was obtained by linearizing the shift-and-inverted QEP written in Eq. (6.6). We may also reverse the order of the linearization and shift-and-invert operations. Performing a linearization of Eq. (3.56) that results in an eigenproblem $\widehat{\mathbf{A}}\mathbf{x} = \lambda\mathbf{x}$ of double size, and subsequently a shift-and-invert transformation arriving at $(\widehat{\mathbf{A}} - \sigma\boldsymbol{I})^{-1}\mathbf{x} = \mu\mathbf{x}$, shows that the matrix applied in the Arnoldi procedures can also be written [63]

$$(\widehat{\mathbf{A}} - \sigma\boldsymbol{I})^{-1} = \begin{pmatrix} -\mathbf{M}^{-1}\widehat{\mathbf{C}} & -\mathbf{M}^{-1}\mathbf{K} \\ \boldsymbol{I} - \sigma\mathbf{M}^{-1}\widehat{\mathbf{C}} & -\sigma\mathbf{M}^{-1}\mathbf{K} \end{pmatrix}, \tag{6.13}$$

where

$$\widehat{\mathbf{C}} = \bar{\mathbf{H}}_L + \sigma\bar{\mathbf{H}}_{L,L}. \tag{6.14}$$

The eigenpairs $(\mu_k, \mathbf{x}_k)$ of $(\widehat{\mathbf{A}} - \sigma\boldsymbol{I})^{-1}\mathbf{x} = \mu\mathbf{x}$ are exactly the same as of Eq. (3.58). In addition, we may now consider the combined spectral transformation for two conjugate shifts σ and σ^*, given by

$$\mathbf{T} = (\widehat{\mathbf{A}} - \sigma\boldsymbol{I})^{-1}(\widehat{\mathbf{A}} - \sigma^*\boldsymbol{I})^{-1} = \frac{\mathrm{Im}\{(\widehat{\mathbf{A}} - \sigma\boldsymbol{I})^{-1}\}}{\mathrm{Im}\{\sigma\}}, \tag{6.15}$$

which was originally proposed by Parlett and Saad [84]. Applying matrix $\mathbf{T}$ in the Arnoldi procedure generates approximate solutions to $\mathbf{T}\mathbf{x} = \mu'\mathbf{x}$, where the eigenvalues are given by

$$\mu' = \frac{1}{(\lambda - \sigma)(\lambda - \sigma^*)}, \tag{6.16}$$

which become extreme for conjugate eigenvalues λ and λ^* of Eq. (3.56) that are close to σ and σ^*. In our case, the complex shifts are purely imaginary: $\sigma = \hat{\imath}\beta$, where β is real. Then we have $\mu' = (\lambda^2 + \beta^2)^{-1}$ and, more importantly, the matrix $\mathbf{T}$ is simply given by β^{-1} times the imaginary part of Eq. (6.13), written as

$$\mathbf{T} = \begin{pmatrix} -\beta^{-1}\mathrm{Im}\{\mathbf{M}^{-1}\widehat{\mathbf{C}}\} & -\beta^{-1}\mathrm{Im}\{\mathbf{M}^{-1}\mathbf{K}\} \\ \mathrm{Re}\{\mathbf{M}^{-1}\widehat{\mathbf{C}}\} & \mathrm{Re}\{\mathbf{M}^{-1}\mathbf{K}\} \end{pmatrix}, \tag{6.17}$$

which is purely real. This makes it feasible to use real arithmetic in all parts of the algorithm except for the initial complex LU-factorization of $\mathbf{M}$.

6.2.2 Selection scheme and convergence criteria

In order to benefit from the shift-and-invert strategy described above we must separate the solutions of the QEPs into three groups, one for each shift. This is illustrated schematically in Fig. 6.1, where the unit disc of the complex plane is cut into "quarter-slices". In practice, the separation is accomplished by an simple selection scheme to determine which of the available solutions $(\lambda_k, \mathbf{c}_k)$ that correspond to wanted Ritz pairs located inside the valid quarter-slice. The selection scheme can be implemented as two separate processes.

The first selection process is designed to identify those solutions that correspond to eigenpairs of the original QEP which satisfy $\lambda_{\min} \leq |\lambda_k| \leq 1$. It is important to realize, however, that since all computations are done in finite precision arithmetic, there is no guarantee that the Bloch modes of the electrodes will have magnitudes $|\lambda_k|$ exactly equal to 1. Even the left-going propagating modes, which are targeted in our case, can have $|\lambda_k| > 1$. In practice, we therefore define the propagating modes to be those Ritz pairs $(\lambda_k, \mathbf{c}_k)$ that satisfy

$$(1 + \epsilon)^{-1} \leq |\lambda_k| \leq 1 + \epsilon \tag{6.18}$$

where ϵ is a small infinitesimal (set to 10^{-8} in our implementation). In order to make sure that all propagating modes are taken into consideration it is thus necessary to select all Ritz pairs that satisfy $\lambda_{\min} \leq |\lambda_k| \leq 1 + \epsilon$.

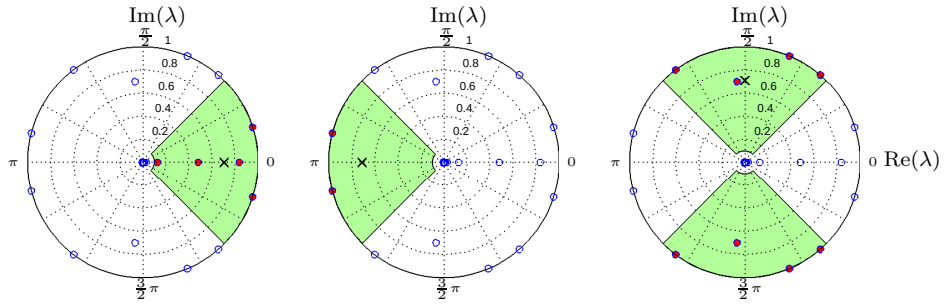


Figure 6.1: Illustration of the complex eigenvalues (blue/circles) for the Al(100) electrode at $E = 3$ eV. The eigenvalues corresponding to the right-going modes (red/filled dots) of interest can be separated according to their location within three distinct (green/shaded) areas of the unit disc and determined efficiently using shift-and-invert spectral transformations to $\pm 1/\sqrt{2}$ and $\hat{\imath}/\sqrt{2}$ (crosses).

To obtain the Ritz values λ_k used in the selection process, we have to transform the solutions of the projected eigenproblem to the corresponding Ritz pairs $(\lambda_k, \mathbf{c}_k)$ by reversing the shift-and-invert operation. The transformation again depends on whether the shift σ is real or imaginary. In the case of real σ, we have $\lambda_k = \mu_k^{-1} + \sigma$ from Eq. (6.10). For imaginary σ, Eq. (6.16) can be rearranged to $\lambda_k^2 = \mu_k^{-1} + \sigma^2$, which has two solutions of equal magnitude.

This is sufficient to allow selection on the basis of the magnitude $|\lambda_k|$, however, when it comes to obtaining the Ritz values λ_k themselves, it is necessary to use other means for imaginary σ, e.g., by forming the Rayleigh quotient [83]. In our case, and for QEPs in particular, it is possible and computationally advantageous to use alternatives to the Rayleigh quotient that work with vectors and matrices of size m_L instead of $2m_L$. Several such techniques that are both fast and accurate have recently been devised by Hochstenbach and van der Vorst [85]. We will adopt the MR-2 method of that paper, which yields $\lambda_k = \frac{\alpha_k}{\beta_k}$, for α_k and β_k defined as

$$\begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix} = -\tilde{\boldsymbol{Z}} \bar{\mathbf{H}}_{L,L}^{\dagger} \mathbf{c}_k, \tag{6.19}$$

where $\tilde{\boldsymbol{Z}}$ is the pseudo-inverse of $\boldsymbol{Z} = (\bar{\mathbf{H}}_{L,L}\mathbf{c}_k, \bar{\mathbf{H}}_L\mathbf{c}_k)$. Since all eigenvectors are unchanged by the shift-and-invert operation, the $\mathbf{c}_k$ vectors applied here are the first m_L elements of the Ritz vectors.

The remaining selection process should single out the Ritz pairs that are inside the valid slice of the unit disc. To this end, we can apply the inner product of $(\mathrm{Re}\{\lambda_k\}, \mathrm{Im}\{\lambda_k\})$ and $(\mathrm{Re}\{\sigma\}, \mathrm{Im}\{\sigma\})$, given by

$$\mathrm{Re}\{\lambda_k\}\mathrm{Re}\{\sigma\} + \mathrm{Im}\{\lambda_k\}\mathrm{Im}\{\sigma\} = |\lambda_k||\sigma|\cos\theta \tag{6.20}$$

where θ is the angle between λ_k and σ in a polar representation of the complex plane. In order for λ_k to be inside the quarter-slice that has σ on the bisector we must have $|\theta| \leq \pi/4$ or equivalently $\cos\theta \geq 1/\sqrt{2}$. For real shifts $\sigma = \pm 1/\sqrt{2}$, this observation yields the condition

$$\frac{\mathrm{Re}\{\lambda_k\}\mathrm{Re}\{\sigma\}}{|\lambda_k|} \geq \frac{1}{2}, \tag{6.21}$$

and similarly for imaginary shift $\sigma = \hat{\imath}/\sqrt{2}$,

$$\frac{|\mathrm{Im}\{\lambda_k\}\mathrm{Im}\{\sigma\}|}{|\lambda_k|} > \frac{1}{2}, \tag{6.22}$$

where the absolute value of the left-hand side is taken to allow λ_k to be both in the top and the bottom quarter-slice. Notice that the equality is removed since the (very rare) event of λ_k lying exactly on the border of two slices is already taken into account in the condition for real σ.

Finally, let us discuss how to check for convergence. For each shift, the convergence condition is regarded as satisfied when all the Ritz pairs of interest that are also located inside the valid quarter-slice are identified and accurate to a given tolerance. We estimate the accuracy of the obtained pairs $(\lambda_k, \mathbf{c}_k)$ by evaluating the corresponding relative residual norm, which yields the following convergence criterion,

$$\frac{||(\bar{\mathbf{H}}_{L,L}^{\dagger} + \lambda_k \bar{\mathbf{H}}_L + \lambda_k^2 \bar{\mathbf{H}}_{L,L})\mathbf{c}_k||_2}{\mathrm{norm}(\bar{\mathbf{H}}_L)} \leq \texttt{tol} \tag{6.23}$$

where `tol` is the convergence tolerance and $\mathrm{norm}(\bar{\mathbf{H}}_L)$ is an appropriate norm for matrix $\bar{\mathbf{H}}_L$. In our implementation we set $\texttt{tol} = 10^{-11}$ and apply the approximation $\mathrm{norm}(\bar{\mathbf{H}}_L) \approx ||\mathrm{diag}(\bar{\mathbf{H}}_L)||_2$, that is, we include only the diagonal entries of the 2-norm of $\bar{\mathbf{H}}_L$. These choices require very low computational effort and give the correct result for all numerical examples we have investigated.

6.2.3 Restarting and multiple eigenvalues

An unfortunate aspect of the Arnoldi procedure is that one cannot know in advance how many steps will be needed before the eigenpairs of interest are well approximated by Ritz pairs. If many steps are necessary, then solving the projected eigenvalue problem becomes costly. Moreover, when applying our Krylov method to evaluate the self-energy matrices we do not know the exact number of eigenpairs wanted and cannot estimate the required dimension of the Krylov subspace.

One way to circumvent the first difficulty is to restart the Arnoldi method after a certain number of iterations using the obtained information to generate a better starting vector or deflate particular eigenvalues [80]. However, this

will not improve on the second difficulty which requires an adaptive maximum dimension of the Krylov subspace. In addition, we observe in most applications that the benefits of an efficient restart procedure (e.g., as devised by Morgan and Zeng [69]), does not outweigh the computational expense of the restarting overhead. The typical size of the self-energy matrices encountered is too small to make it beneficial to use such techniques, which have been developed for large-scale applications.

Therefore, we have chosen to employ a simple continuation scheme instead of restarting, where a check for convergence is performed after a given number of Arnoldi iterations, and if not satisfied, the procedure simply continued where it was left off. With the input parameter k, the listed ALGORITHM XI is able to generate an initial Krylov subspace $\mathcal{K}_m$ of a given dimension m, but also continue the process augmenting the space with subsequent calls. This allows us to perform iterations as long as the approximations are unsatisfactory and/or there is doubt whether all wanted eigenpairs have been found.

An important special case to be considered when applying the Arnoldi procedure to solve an eigenvalue problem is the possibility of multiple eigenvalues. A Krylov subspace method will, in theory, produce only one eigenvector corresponding to a multiple eigenvalue. So determining multiplicity is quite difficult. Several approaches exist that deal with this problem, including deflation combined with effects of round-off error [80], block Arnoldi procedures [80] and so-called random restarts [86, 69]. The present Krylov method does not incorporate any mechanisms to take algebraic multiplicity into account because such cases do not occur in practice for the applications of this work (eigenvalues will not be identical to machine precision in any of the numerical examples, but only to within $\sim 10 - 11$ digits, see Sect. 6.3).

6.2.4 Implementation

The implementation of our Krylov method is composed of two main parts. An iterative part that determines the wanted Ritz pairs $(\lambda_k, \mathbf{c}_k)$ which approximate the eigenpairs of the QEP in Eq. (3.56), and a non-iterative part that sets up the mode and phase matrices and evaluates the self-energy matrix from these by direct methods. The iterative part is organized as three independent computations, one for each of the used shifts σ. It consists of the application of the Arnoldi procedure together with a check for convergence plus the initial work to construct the input matrices for ALGORITHM XI. As described above the actual calculations will depend on whether the shift is real or imaginary.

The key steps of the Krylov method for evaluating the self-energy matrix $\mathbf{\Sigma}_L$ of the left electrode are presented in ALGORITHM XII below. It is important to stress that the details of each step are kept at a minimum to enhance the readability. Furthermore, for evaluating the self-energy matrix $\mathbf{\Sigma}_R$ of the right electrode, the steps are exactly the same, except for the substitution $L \to R$ of all super- and subscripts and the removal of line 1 (this line is only required

for left electrodes in order to obtain $\mathbf{\Sigma}_L$ also from right-going modes). We refer the reader to **PAPER III** for a detailed discussion of the steps in the algorithm.

ALGORITHM XII: *Krylov method to evaluate* $\tilde{\mathbf{\Sigma}}_L$

1. Exchange matrices $\bar{\mathbf{H}}_{L,L}$ and $\bar{\mathbf{H}}_{L,L}^{\dagger}$.
2. **for** $\sigma = 1/\sqrt{2}, -1/\sqrt{2}, \hat{\imath}/\sqrt{2}$ **do**
3. **if** σ is real, calculate $\mathbf{A}$ from Eq. (6.12),
   else calculate $\mathbf{T}$ from Eq. (6.17) and set $\mathbf{A} = \mathbf{T}$.
4. Select random vector $\mathbf{v}_1$ of size $2m_L$.
5. Apply ALGORITHM XI to generate $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$.
6. Solve the projected eigenproblem $\mathbf{H}_m\mathbf{y} = \mu_k\mathbf{y}$.
7. **if** σ is real,
   select all $(\mu_k, \mathbf{y}_k)$ that satisfy $\lambda_{\min} \leq |\mu_k^{-1} + \sigma| \leq 1 + \epsilon$, and store the Ritz pairs $(\lambda_k, \mathbf{c}_k) = (\mu_k^{-1} + \sigma, \mathbf{V}_m\mathbf{y}_k)$ that have $\mathrm{Re}\{\lambda_k\}\mathrm{Re}\{\sigma\} \geq \frac{|\lambda_k|}{2}$,
   else
   select all $(\mu_k, \mathbf{y}_k)$ that satisfy $\lambda_{\min} \leq |\mu_k^{-1} + \sigma^2|^{\frac{1}{2}} \leq 1 + \epsilon$, and evaluate the original eigenvalues λ_k with the MR-2 method of Ref. [85] and store the Ritz pairs $(\lambda_k, \mathbf{c}_k) = (\lambda_k, \mathbf{V}_m\mathbf{y}_k)$ that have $|\mathrm{Im}\{\lambda_k\}\mathrm{Im}\{\sigma\}| > \frac{|\lambda_k|}{2}$.
8. For the selected pairs $(\lambda_k, \mathbf{c}_k)$ find residual $||(\bar{\mathbf{H}}_{L,L}^{\dagger} + \lambda_k\bar{\mathbf{H}}_L + \lambda_k^2\bar{\mathbf{H}}_{L,L})\mathbf{c}_k||_2$, and check for convergence. If not satisfied, increase m appropriately and go to step 5.
9. **end**
10. For all stored Ritz pairs $(\lambda_k, \mathbf{c}_k)$ having $(1+\epsilon)^{-1} \leq \lambda_k \leq 1+\epsilon$, calculate velocity v from Eq. (3.61). Discard the pairs with $v < 0$ (i.e, the left-going modes).
11. Construct matrices $\tilde{\mathbf{\Lambda}}_L^+$ and $\tilde{\mathbf{C}}_L^+$ (see Eqs. (3.62) and (3.63)) from the remaining pairs.
12. $[\tilde{\mathbf{Q}}_L^+, \tilde{\mathbf{R}}_L^+] = \mathrm{QR}\{\tilde{\mathbf{C}}_L^+\}$, solve $\tilde{\mathbf{R}}_L^+\tilde{\mathbf{C}}_L^+ = (\tilde{\mathbf{Q}}_L^-)^{\dagger}$
13. $\tilde{\mathbf{B}}_L^+ := \tilde{\mathbf{C}}_L^+\tilde{\mathbf{\Lambda}}_L^+\tilde{\mathbf{C}}_L^+$
14. $\tilde{\mathbf{\Sigma}}_L := \bar{\mathbf{H}}_{L,L}^{\dagger}[\bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}^{\dagger}(\tilde{\mathbf{B}}_L^+)^{-1}]^{-1}\bar{\mathbf{H}}_{L,L}$

6.2.5 Generalization to complex Hamiltonians

In the Krylov subspace method presented above we have assumed that the electrode Hamiltonian matrices are real in order to simplify the computational procedures. We now discuss the steps required to handle the case of com-

plex $\bar{\mathbf{H}}_L$ and $\bar{\mathbf{H}}_{L,L}$, which is the case, e.g., when applying **k**-point sampling (ALGORITHM XII only works for the Γ-point).

As noted in Sec. 6.2.1 the assumption of real $\bar{\mathbf{H}}_L$ and $\bar{\mathbf{H}}_{L,L}$ leads to simplifications with the shift-and-invert operations: Firstly, we may consider only right-going modes $(\lambda_k, \mathbf{c}_k)$ with $|\lambda_k| \leq 1$ since the left-going are uniquely related as $(\lambda_k^{-1}, \mathbf{c}_k)$, and, secondly, we can use the spectral transformation $\mathbf{T}$ in Eq. (6.17) to determine the wanted eigenpairs for the two imaginary shifts $\sigma = \pm\hat{\imath}/\sqrt{2}$ simultaneously and in real arithmetic.

In order to generalize the Krylov subspace method to complex Hamiltonian matrices, it is thus necessary to determine the left-going modes satisfying $1 \leq |\lambda_k| \leq \lambda_{\min}^{-1}$ (i.e, located outside the unit circle) directly, since there is no general relation to the right-going modes (we note that it is advantageous to change the shift positions to outside the unit circle, although this is not necessary for good convergence). Furthermore, we must abandon the $\mathbf{T}$ matrix and perform two independent shift-and-invert operations for $\sigma = \pm\hat{\imath}/\sqrt{2}$. It is clear, that all this is now done in complex arithmetic and that the extra shift required will make the general algorithm a little more expensive (as shown in Sec. 6.3.3, the LU-factorizations required for each shift-and-invert operation is the dominant cost of our approach). We have implemented the generalizations and refer the reader to the appendix of **PAPER III** for a numerical example.

6.3 Convergence behavior and computational complexity

The Krylov method described in Sect. 6.2 represents an iterative approach for evaluating the self-energy matrices. Accuracy and efficiency of the method depends on the dimensions of the three Krylov subspaces generated for the three applied shifts σ. In practice, these dimensions are controlled by two parameters: `tol` and $\lambda_{\min}$. Typically, the smaller `tol` is, the more work is required to satisfy the convergence criteria. However, setting this value too large may cause eigenvalues to be missed when there are nearly multiple or clustered eigenvalues. The $\lambda_{\min}$ parameter on the other hand, sets the scale of exactly how many Ritz pairs will be approximated. In this section, we will exemplify the convergence behavior of ALGORITHM XII by monitoring the relative residual norm in relation to the two controlling parameters `tol` and $\lambda_{\min}$. This gives an indication of the typical number of iterations required for a given size of the self-energy matrix. We also estimate the number of floating point operations needed in the expensive parts of the algorithm and give the computational complexity.

6.3.1 Convergence of the residual norm

In order to demonstrate the convergence behavior of our Krylov method, we first monitor the relative residual norm of the wanted eigenpairs as a function

of the number of iterations. An expression for this norm for a given eigenpair $(\lambda_k, \mathbf{c}_k)$ is available as the left hand side of Eq. (6.23). We will consider the Al(100) electrode at $E = 3$ eV and parameter $\lambda_{\min} = 0.1$, which requires a total of 13 eigenpairs to be determined (8 Bloch modes and 5 evanescent modes) from the three separate Arnoldi procedures. This example thus corresponds to the situation illustrated in Fig. 6.1 and represents a typical calculation for an Al electrode with 18 atoms per unit cell (the size of the self-energy matrix is 72).

In Fig. 6.2 we present curves showing the history of the residual norm for a given number of iterations for the wanted eigenpairs in each of the separate shift-and-invert Arnoldi procedures. We show only the 45 first iterations since this number is enough for convergence in all cases. Also, only residuals for eigenpairs corresponding to right-going modes are displayed.

The top figure of Fig. 6.2 illustrates the results from applying the shift $\sigma = 1/\sqrt{2}$ and shows that the Arnoldi procedure determines four different Ritz pairs with individual convergence curves. Comparing with the respective polar plot in Fig. 6.1 (top-left), we observe a fifth eigenvalue ($\lambda = 0.95 + 0.31\hat{\imath}$) located inside the valid quarter-slice. This fifth eigenvalue represents a left-going mode and is thus discarded in step 10 of ALGORITHM XII. We also see by comparison with Fig. 6.1, that the eigenpairs with eigenvalues furthest from the current shift (the cross) in the complex plane, in this case λ_4, is the slowest to converge. This is characteristic of how the shift-and-invert Arnoldi method locates eigenvalues [83], and also the key to the success of the proposed Krylov method.

The middle figure of Fig. 6.2 shows the convergence of the two Ritz pairs which are covered by the Arnoldi procedure with $\sigma = -1/\sqrt{2}$ and correspond to right-going modes in the present example. We note that λ_5 and λ_6 are nearly multiple eigenvalues and that the behavior of the residual norms, where one eigenpair is available many iterations before its counterpart, is typical in such a case. Here, in particular, we see that eigenvalue λ_5 is determined to an accuracy of $\sim 10^{-11}$ after 18 iterations before λ_6 even shows up as a Ritz value of the projected eigenproblem. This indicates that λ_5 and λ_6 must be identical to around 10 significant digits, and that they cannot be distinguished in our Arnoldi procedure before this accuracy is achieved. Again we would like to emphasize that without additional mechanisms to deal with multiple eigenvalues, this implies an upper bound condition on the value of the `tol` parameter. Thus if all multiple eigenvalues are to be identified, which is necessary for ALGORITHM XII to produce the correct result, then this parameter should not be larger than the absolute difference between any two of the wanted eigenvalues.

The bottom figure of Fig. 6.2 shows the residual norm history of the remaining 7 Ritz pairs required in the current example. These are determined by the Arnoldi procedure with imaginary shift $\sigma = \hat{\imath}/\sqrt{2}$ and correspond to filled dots in the bottom polar plot of Fig. 6.1 that represent right-going modes. We observe that the eigenvalue closest to σ, here denoted by λ_8, constitutes a complex conjugate pair together with λ_9, and that these have exactly the same residual norm curve (indistinguishable in the figure) although they are obtained

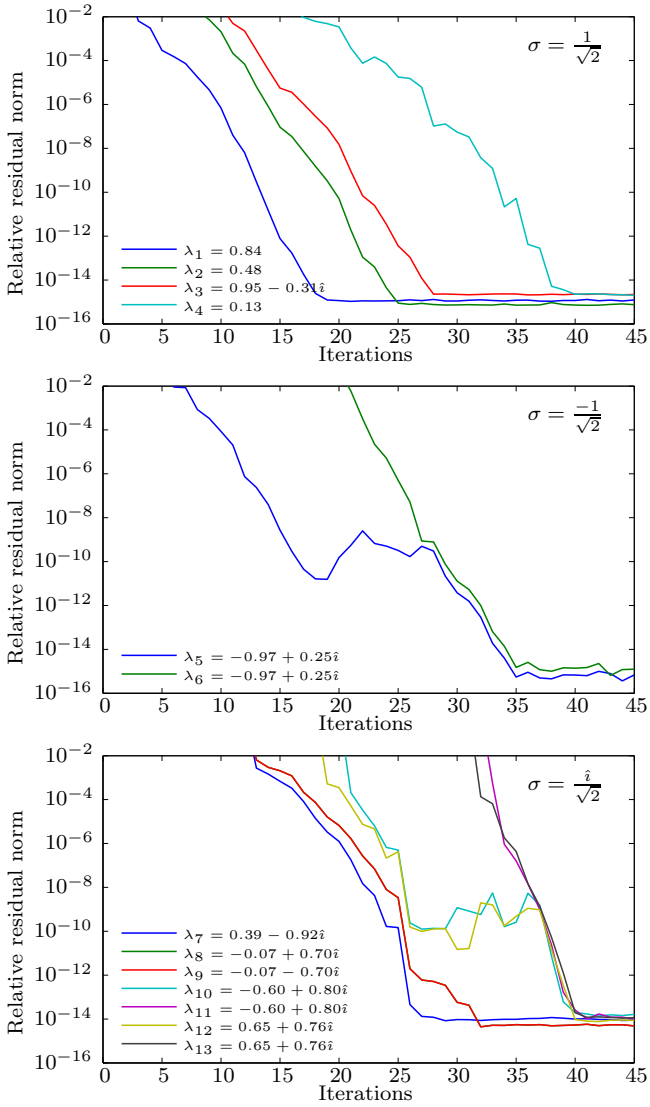


Figure 6.2: Convergence behavior of the Krylov algorithm for the Al(100) electrode at $E = 3$ eV. The figures show the residual norm as a function of iterations for Ritz pairs that satisfy $0.1 \leq |\lambda_k| \leq 1 + \epsilon$, in the case of shift-and-invert transformations to $\pm 1/\sqrt{2}$ and $\hat{\imath}/\sqrt{2}$, respectively.

separately as individual Ritz pairs in the algorithm. The reason for the identical residual norm is that we apply the combined spectral transformation $\mathbf{T}$ given by Eq. (6.15) in the case of the imaginary shift. We also note that there are two more instances of nearly multiple eigenvalues (λ_{10},λ_{11} and λ_{12},λ_{13}) which also have the characteristic residual norm curves of almost identical Ritz values.

In all residual norm figures, we see the trend that the eigenvalues located far from the position of the shift are slow to converge. This suggests that eigenvalues located in the vicinity of the intersections between the unit circle and the dividing lines of the four quarter-slices will be the most difficult to determine since they are furthest from the respective shifts. The maximum distance from such an eigenvalue to σ is $1/\sqrt{2}$, which is the same as from σ to the origin. This rouses concern whether the many unwanted eigenvalues close to the origin can become dominant compared to the wanted border eigenvalues. Fortunately, this will not be the case because the unwanted eigenvalues close to the origin are clustered and therefore easy to represent in the Krylov subspace with only a few iterations [83]. We observe this in practice, e.g., from the bottom figure of Fig. 6.2, where the Ritz pair corresponding to λ_{12}, which lies close to the worst case position on the unit circle, initially converges only slightly slower than the Ritz pair for λ_8 positioned right next to the shift.

6.3.2 Number of iterations to convergence

The curves of the residual norms presented above indicate that the proposed Krylov method is an iterative procedure that locates the specific eigenpairs required to evaluate the self-energy matrices in a systematic fashion. We will now consider the typical number of iterations performed by the method. This number is obviously related to the number of wanted eigenpairs which again depends on the parameter $\lambda_{\min}$ and the particular energy E. Also the convergence tolerance `tol` may affect the number of iterations, however, since the convergence is always linear and fast in the final part of the Arnoldi procedure (see the steep slopes of the residual curves in Fig. 6.2), this influence is minor. In the following, we therefore fix $\texttt{tol} = 10^{-11}$ and measure the number of iterations required for different energies E and three settings of $\lambda_{\min}$.

First we look at the Al(100) electrode of the previous section. The parameter $\lambda_{\min}$ is set to 0.05, 0.1, and 0.5, which are reasonable values for two-probe systems with this electrode (see, e.g., numerical examples in Sect. 6.4). In the upper panel of Fig. 6.3 we present the number of wanted Ritz pairs determined by the Krylov method when it is applied to the Al(100) electrode at energies in the interval $E \in [-12\text{ eV}, 20\text{ eV}]$ and steps of 0.5 eV between each data point. The results are displayed in a stair-step fashion to indicate that the curve does not alter much if the energy resolution of the measurements is increased. In the lower panel we show the corresponding total number of iterations performed by the Arnoldi procedures. This number is rapidly varying even for very small step size between the energy points since it depends on the specific positions of

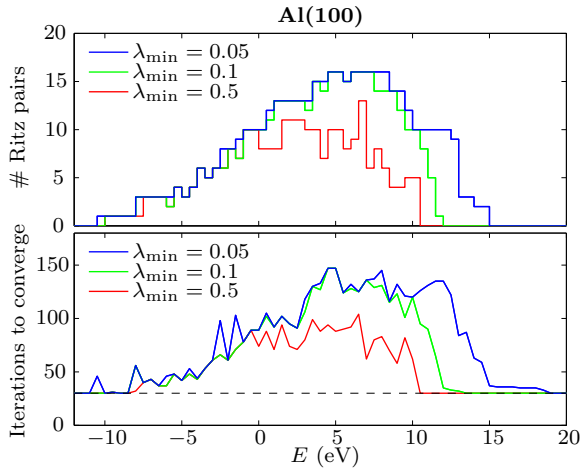


Figure 6.3: Typical number of iterations with the Krylov method for the Al(100) electrode. Upper panel: the number of wanted Ritz pairs to be determined at different energy points for three settings of the parameter $\lambda_{\min}$. Lower panel: the corresponding total number of iterations performed by the three Arnoldi procedures in order to reach convergence for $\texttt{tol} = 10^{-11}$.

the wanted Ritz pairs and how difficult they are to locate. For E outside the energy range shown, there are no propagating modes nor any evanescent modes satisfying $\lambda_{\min} \leq |\lambda_k| < (1+\epsilon)^{-1}$.

From Fig. 6.3 we can see that the total number of iterations in general follows to the number of Ritz pairs required, as expected. The minimum number of iterations for a given E is fixed by the value m provided as input to ALGORITHM XII, which designates the dimension of the initial Krylov subspace. Here we have chosen $m = 15$ which implies a minimum number of 45 iterations and is illustrated by the dashed line in the lower panel (in practice, a value of $m = 30$ is more reasonable, but here we halve this to obtain more details in the curves). The maximum number of iterations measured is 147 at $E = 5$ eV for $\lambda_{\min} = 0.05$ and $\lambda_{\min} = 0.1$, and 104 at $E = 6.5$ eV for $\lambda_{\min} = 0.5$. For all energies, the number of iterations is either stagnant or decreasing as the parameter $\lambda_{\min}$ is increased, since the area of the unit disc for which eigenvalues are taken into account becomes smaller.

In Fig. 6.4 we consider an armchair (4,4) carbon nanotube (CNT) electrode with 16 carbon atoms per unit cell and the corresponding (12,12)–CNT electrode with 48 atoms per unit cell. The sizes of the self-energy matrices in these cases are 128 and 384, respectively. The energy interval investigated is $E \in [-27 \text{ eV}, 27 \text{ eV}]$ with steps of 0.5 eV between points. When E is outside this

interval the result is again zero wanted Ritz pairs.

It is evident from Fig. 6.4 that the overall trend for the CNT electrodes is the same as for Al(100). The more Ritz pairs to be determined the more iterations are required for convergence. We note, however, that the amount of Ritz pairs needed for the (4,4)–CNT electrode, and hence the number of iterations to converge, is on the same level as for the Al(100) electrode in Fig. 6.3, even though the sizes of the matrices entering the calculation are almost doubled. Thus the number of Ritz pairs required for a given problem size is relatively smaller, which means that a CNT electrode is an easier task for the Krylov method compared to the Al(100) electrode.

We also note that the number of Ritz pairs for a given E and $\lambda_{\min}$ in the case of the (12,12)–CNT electrode is always close to three times the corresponding number in the case of the (4,4)–CNT electrode, and likewise for the total number of iterations. E.g., when $\lambda_{\min} = 0.05$, the maximum number of iterations is 162 for the (4,4)–CNT electrode and 404 for the (12,12)–CNT electrode. This is not surprising since there are exactly three times as many right-going modes in the (12,12)–CNT electrode than in the (4,4)–CNT electrode. In other words, the efficiency of the proposed Krylov method for the large electrode is similar to that achieved for the small electrode.

The two observations made from Fig. 6.4 indicate that the total number of iterations does not depend directly on the size N of the matrices at hand, but only on the specific number of right-going modes designated as wanted via the parameter $\lambda_{\min}$. This is an important feature, since it means, that the three Arnoldi procedures, in practice, have an operation cost which is low compared to other parts of the Krylov method, as shown explicitly in the following and by numerical examples in Sec. 6.4.

6.3.3 Computational complexity

In this section we discuss the main computational expenses of ALGORITHM XII. Estimates of floating point operations for the key steps are presented in Table 6.1. The expressions listed are based on the assumption that the number of wanted Ritz pairs to be determined is q, and that this requires a total of r iterations in the Arnoldi procedure. The size of the input matrices is N.

In the first two lines of the table, we have estimated the cost of the initial calculation of the matrices $\mathbf{A}$ and $\mathbf{T}$ that incorporate the shift-and-invert tranformation and the linearization of the QEP to be solved. As shown in Sect. 6.2.1, this calculation can be accomplished by a single LU-factorization of $\mathbf{M}$ in Eq. (6.7) and four substitution procedures to obtain $\mathbf{M}^{-1}\mathbf{K}$ and $\mathbf{M}^{-1}\mathbf{C}$. The LU-factorization is known to require $\frac{2}{3}N^3 + O(N^2)$ operations for a matrix of size N [1] and the subsequent substitutions will fall under the second term of this expression. In the case of an imaginary shift, however, the LU-factorization and substitutions have to be done in complex arithmetic. We then assume an operations count that is (up to) 6 times higher (also in the following).

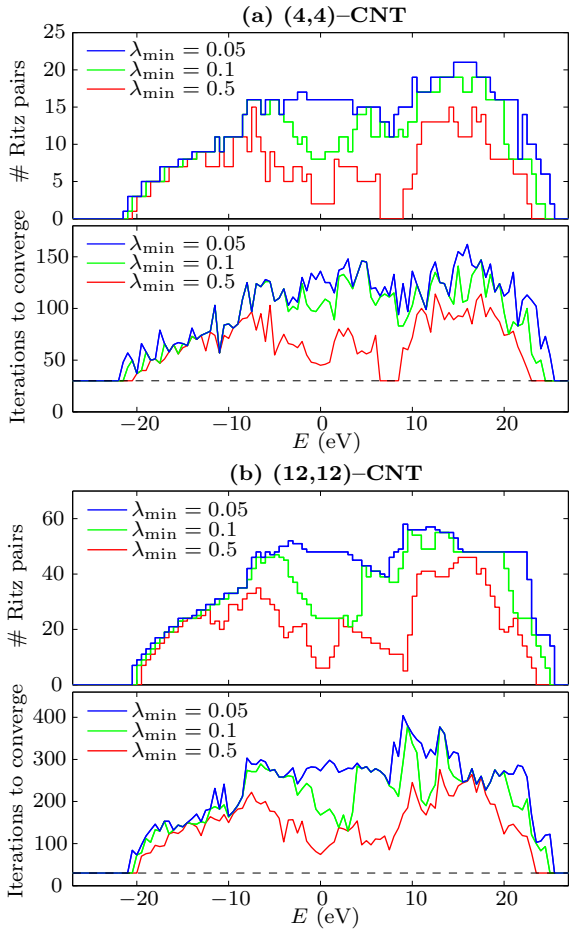


Figure 6.4: Typical number of iterations with the Krylov method for (a) the (4,4)–CNT electrode and (b) the (12,12)–CNT electrode. See the description given in the caption of Fig. 6.3 for more details.

Table 6.1: Estimated number of floating point operations required in the main steps of the Krylov method when applied to matrices of size N. We assume that r is the final number of iterations required by the Arnoldi procedure to determine exactly q wanted Ritz pairs.

Step in the Krylov method	Operations estimates
Calculate $\mathbf{A}$ from Eq. (6.12)	$\sim \frac{2}{3}N^3 + O(N^2)$
Calculate $\mathbf{T}$ from Eq. (6.17)	$\sim \frac{12}{3}N^3 + O(N^2)$
Arnoldi procedure, $\sigma = \pm\frac{1}{\sqrt{2}}$	$\sim 4N^2r + 4Nr^2 + O(r)$
Arnoldi procedure, $\sigma = \frac{i}{\sqrt{2}}$	$\sim 8N^2r + 4Nr^2 + O(r)$
Solving $\mathbf{H}_k\mathbf{y} = \mu\mathbf{y}$	$\sim O(r^3)$
Selection of Ritz pairs $(\lambda_k, \mathbf{c}_k)$	$\sim 4Nr + O(r)$
Computing residual norms	$\sim 12N^2q + O(Nq)$
$\tilde{\mathbf{B}}_L^+ := \tilde{\mathbf{C}}_L^+\tilde{\mathbf{\Lambda}}_L^+\tilde{\mathbf{C}}_L^+$	$\sim 12N^2q + O(Nq^2)$
$\tilde{\mathbf{\Sigma}}_L := \bar{\mathbf{H}}_{L,L}^\dagger[\bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}\tilde{\mathbf{B}}_L^+]^{-1}\bar{\mathbf{H}}_{L,L}$	$\sim (3\times 4 + \frac{12}{3})N^3 + O(N^2)$

For the Arnoldi procedure given in ALGORITHM I the computational expense of one iteration is dominated by a matrix-vector product and the orthogonalization loop. The matrix-vector product costs $8N^2$ operations for the square matrix $\mathbf{A}$ of size $2N$. For real shifts, however, the top half of $\mathbf{A}$ in Eq. (6.12) is given by one zero block and one identity block, that does not have to be multiplied. In the rth iteration the Gram-Schmidt orthogonalization takes $2Nr$ operations. For a total of r iterations this then costs $2Nr^2$ operations. Unfortunately, inexact arithmetic can cause the orthogonalization step to fail producing orthogonal vectors [83], in which case, reorthogonalization is required. Taking this into account by multiplying the orthogonalization expense by 2, we estimate the worst-case computational cost of the Arnoldi procedure for real shifts to be $4N^2 + 4Nr^2$ plus less significant terms that can be collected as $O(r)$. For imaginary shift, the full size $2N$ matrix-vector multiplication is necessary which doubles the factor on the first term.

Every time a check for convergence is performed between the Arnoldi procedures, it is necessary to solve the standard eigenvalue problem $\mathbf{H}_k\mathbf{y} = \mu\mathbf{y}$, which corresponds to the QEP projected onto the available Krylov subspace. This is done by applying a direct method (i.e., DGEEV from the LAPACK library) that takes $O(r^3)$ operations when the subspace has dimension r (in order to save some of the computational work, it is possible to exploit that the projected matrix $\mathbf{H}_k$ produced in the Arnoldi procedures is in upper Hessenberg form). We also note that as long as $r \ll N$, the cost of $O(r^3)$ operations will be a less significant expense.

The selection of the wanted Ritz pairs in step 7 of ALGORITHM XII is implemented by simply running through the $2k$ solutions of the projected eigenprob-

lem and discarding those Ritz pairs $(\lambda_k, \mathbf{c}_k)$ that do not satisfy the conditions. The most work is then used in obtaining the Ritz vectors $\mathbf{c}_k = \mathbf{V}\mathbf{y}_k$, which requires $4Nr$ operations for the r complex vectors $\mathbf{y}_k$ that correspond to right-going modes. Assuming that exactly q Ritz pairs survives the selection process, the residual norms of these are computed, and this requires three size N matrix-vector products each plus less expensive inner products. Again, the vectors are complex bringing the operations estimate to $12N^2q + O(Nq)$ for obtaining the residual norms.

To compute the pseudo-inverse is we perform a QR factorization of $\tilde{\mathbf{C}}_L^+$ (assuming that this has full rank) to obtain $\tilde{\mathbf{Q}}_L^+$ and $\tilde{\mathbf{R}}_L^+$, and then solving $\tilde{\mathbf{R}}_L^+\tilde{\mathbf{C}}_L^+ = (\tilde{\mathbf{Q}}_L^+)^\dagger$, where $\tilde{\mathbf{R}}_L^+$ is upper triangular. In this case, the complex ("skinny") QR factorization costs $12Nq^2 + O(q^3)$ operations [1], which is much more than the expense of the subsequent back-substitution and the same as the $12N^2q$ of the matrix-matrix multiplication performed in $\tilde{\mathbf{B}}_L^+ := \tilde{\mathbf{C}}_L^+\tilde{\mathbf{\Lambda}}_L^+\tilde{\mathbf{C}}_L^+$ afterwards. Again we note that for $q \ll N$, this step of the algorithm does not represent a heavy cost.

As can be seen from Table 6.1, the most expensive part of the Krylov method for large N and $r \ll N$, is the initial LU-factorizations of $\mathbf{M}$ needed for matrices $\mathbf{A}$ and $\mathbf{T}$, and the final evaluation of $\tilde{\mathbf{\Sigma}}_L$ (we have to do three complex-matrix-real-matrix multiplication for size $N \times N$ and one complex LU-factorization, which has the unavoidable cost of $(3 \times 4 + \frac{12}{3})N^3$ operations). These steps have an $O(N^3)$ computational complexity. In contrast, the actual iterative part of ALGORITHM XII is dominated by the matrix-vector products that costs $O(N^2)$ operations for every iteration. We showed in the previous section, that the total number of iterations r is related to N, but can be assumed to satisfy $r \ll N$ in most cases. Thus the cost of the Arnoldi iterations applied in the proposed Krylov method will be negligible compared to the other parts of the algorithm for most electrode sizes.

6.4 Applications

In order to illustrate the accuracy and practical aspects of the proposed Krylov subspace method we now present transmission and current calculations for a selection of nano-scale systems. We begin by considering the two example systems Au–DTB–Au and Al–C×7–Al which have been widely studied in the literature so that results can be easily verified. We compute the current through these systems at 1 V and 2 V biases, and use the parameter $\lambda_{\min}$ to investigate the significance of the evanescent modes in obtaining the correct currents. Next, we look at a large carbon nanotube field-effect transistor (CNFET) that displays so-called band-to-band tunneling. Last, we apply the method to evaluate the self-energy matrices of a variety of electrodes (different types and sizes) and compare the actual measured CPU times with those required by conventional methods.

6.4.1 Benzene di-thiol molecule coupled to gold electrodes

Electron transport through molecules attached to metallic electrodes has attracted much interest in recent years. One common example is the case of a benzene di-thiol (DTB) molecule coupled to gold (111) surfaces which has been examined both experimentally [87, 88] and theoretically [45, 41, 46], and has become the closest to a *de facto* benchmark system for numerical electron transport methods (although there is no general agreement among theoretical and experimental results at present). We will apply our Krylov subspace method in a calculation of its transport characteristics and attempt to reproduce the results obtained by other groups for a similar setup.

The geometry of the Au–DTB–Au system we consider here is illustrated in Fig. 3.8. In this system, the Au(111) electrode unit cell contains 27 atoms positioned as three adjacent 3×3 layers. The central region consists of three Au(111)–(3×3) layers, the DTB, and another two Au(111)–(3×3) layers, yielding a structure that has mirror symmetry but where the right and left Au(111) electrode unit cells are not identical. We assume thiolate bonds between the sulfur end-groups and the gold surfaces and use 2.39 Å for the Au–S distance and 1.75 Å for the S–C distance (which makes this setup the same as studied by Stokbro *et al* in Ref. [41]).

We apply the proposed Krylov subspace method to calculate the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ of the left and right electrodes for a range of energies $E \in [-4 \text{ eV}, 4 \text{ eV}]$ and for different choices of the parameter $\lambda_{\min}$. The self-energy matrices are then used in the evaluation of the corresponding transmission coefficients $T(E)$.

In Fig. 6.5 we present the obtained transmission curves in three cases of the bias $V_b = 0$ V, 1 V and 2 V (the Fermi energy E_F is set to 0 eV). The transmission spectra calculated with $\lambda_{\min} = 0.01$ (black/full curves) represent close to exact results. These curves are determined from self-energy matrices that are evaluated by taking both the propagating modes and almost all the evanescent modes into account. By comparison it is apparent that the curve for case $V_b = 0$ V is in complete agreement with that obtained in Ref. [41] and also quite similar to the results in other papers [45, 46].

The (red) dashed curves in Fig. 6.5 correspond to the choice $\lambda_{\min} = 0.5$ and thus depict the $T(E)$ obtained when only the propagating modes and a few evanescent modes (those satisfying $0.5 < |\lambda_k| < 1$) are included in the evaluation of $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$. We see that these spectra are almost indistinguishable from the $\lambda_{\min} = 0.01$ results. This indicates that the majority of the evanescent modes are insignificant for transmission calculations of the Au–DTB–Au system. Furthermore, the setting $\lambda_{\min} = 0.99$ yields self-energy matrices that only take propagating modes or modes very close to propagating into account. Also for this setting, the agreement of the resulting (blue) dotted transmission spectra with those for the other choices of $\lambda_{\min}$ is almost perfect except at a few distinct values of E.

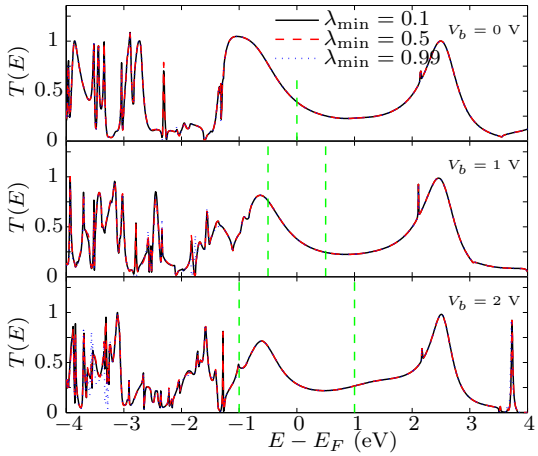


Figure 6.5: Transmission spectrum of the Au–DTB–Au system for different bias voltages V_b. The self-energy matrices used in the $T(E)$ calculations have been obtained by the proposed Krylov subspace method with parameter $\lambda_{\min}$ at several settings: 0.01 (black/full curve), 0.1 (red/dashed curve) and 0.99 (blue/dotted curve). The bias windows are indicated by the vertical dashed lines.

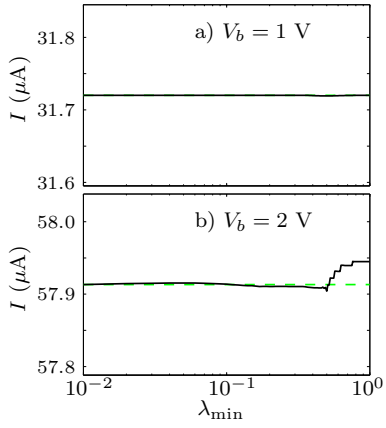


Figure 6.6: Current as a function of the parameter $\lambda_{\min}$ used by the Krylov subspace method for the Au–DTB–Au system with applied bias voltages (a) $V_b = 1$ V and (b) $V_b = 2$ V. The correct currents obtained by conventional methods are $I \approx 31.7$ μA and $I \approx 57.9$ μA, respectively, indicated here by the green/dashed lines.

We can obtain an implicit estimate of the influence of the evanescent modes on the transmission spectrum by calculating the current I as a function of $\lambda_{\min}$ for bias 1V and 2V. Each current calculation involves the integral over the bias windows indicated in Fig. 6.5 of an integrand proportional to $T(E)$ as described in Sec. 3.1.3. Since the curves for $T(E)$ in the bias windows are almost identical for all three choices of $\lambda_{\min}$ we would expect the dependence of this parameter on the current to be minor. As seen in Fig. 6.6, where we show the results from monitoring the current while increasing $\lambda_{\min}$ from 0.01 all the way to 1, this is also the case. Even though there are indications for $V_b = 2$ V that the computed I deviates from the correct value (the dashed lines) when less and less evanescent modes are take into account in evaluating the self-energy matrices, we see that the error is always less than 0.1% and therefore of no real significance.

6.4.2 Carbon wire between aluminum electrodes

To further illustrate the applicability of the proposed Krylov subspace method we also consider carbon chains coupled to metallic electrodes, which have been investigated in detail recently [65, 47, 39]. Carbon atomic wires are interesting conductors since the equilibrium conductance of short mono-atomic chains varies with their length in a oscillatory fashion. It has been shown that the coupling of the wire to the metal electrodes leads to significant charge-transfer doping of the wire [65]. This charge-transfer is facilitated in our formalism via the self-energy matrices, which makes it a well suited test example.

We will examine the Al–C×7–Al two-probe system shown in Fig. 3.8 corresponding to a straight wire of seven carbon atoms attached to Al(100) electrodes. This structure exhibits a local maximum in the oscillatory conductance since it represents an odd-numbered C chain [65]. In our setup, the Al(100) electrode unit cell consists of 18 atoms in four layers with identical unit cells for the left and right electrodes. The same system is studied by Brandbyge *et al* [39].

Again we apply the Krylov subspace method to calculate the self-energy matrices $\mathbf{\Sigma}_L$ $(= \mathbf{\Sigma}_R)$ of the electrodes for energies $E \in [-4 \text{ eV}, 4 \text{ eV}]$ and for different choices of the parameter $\lambda_{\min}$ in order to obtain the transmission coefficients. Fig. 6.7 presents the results for bias voltages $V_b = 0$ V, 1 V and 2 V. The (black) full curves corresponding to $\lambda_{\min} = 0.01$ reproduces the transmission spectra obtained in Ref. [39] (for 0 V and 1 V) exactly except for the peak at $E = 3.63$ eV (for 0 V), which is probably due to finer sampling in our work.

We also see in Fig. 6.7 that the curves for the parameter $\lambda_{\min}$ set to 0.01 and 0.5 are almost identical, which indicates that the vast majority of the evanescent modes (those satisfying $|\lambda_k| < 0.5$) have very little influence on $T(E)$ in the energy regime considered. However, when $\lambda_{\min}$ is set to 0.99 (blue/dotted curves), in which case only modes that can be considered as propagating are included in the evaluation of self-energy matrices, there are several noticeable deviations from the other curves. Also inside the bias windows and especially

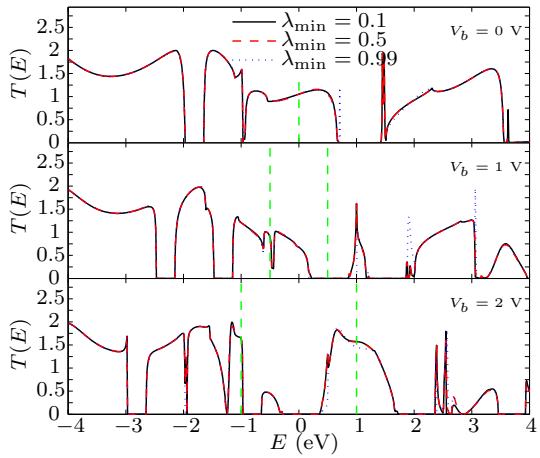


Figure 6.7: Transmission spectrum of the Al–C×7–Al system for different bias voltages V_b. See the caption of Fig. 6.5 for further description.

for $V_b = 2$ V, the disregard of the evanescent modes produces errors in the obtained transmission coefficients $T(E)$.

This effect becomes even more evident in Fig. 6.8, where the current is again displayed as a function of parameter $\lambda_{\min}$ for the non-zero bias voltages. As the value of $\lambda_{\min}$ is increased from around 0.5 to 1, the computed current I starts deviating significantly from the correct value. Therefore, we anticipate that at least the slowly decaying evanescent modes must be taken into account in order to describe the transmission properties of the Al–C×7–Al system. Moreover, we see that this can be achieved in a rigorous and systematic fashion by selecting λ_{min} appropriately when using the proposed Krylov subspace method to calculate the self-energy matrices.

6.4.3 Carbon nanotube field-effect transistor

In this section we will apply the developed method to a nano-device consisting of a CNT stretched between to two metal electrodes and controlled by three gates. The setup is inspired by Appenzeller *et al.* [89], and we expect this particular arrangement to be able to display so-called band-to-band (BTB) tunneling, where one observes gate induced tunneling from the valence band into the conduction band of a semi-conducting CNT and vice versa.

We show the configuration of the band-to-band tunneling two-probe system in Fig. 6.9. The device configuration contains 10 principal layers of a CNT(8,4), having 112 atoms in each. The diameter of the tube and layer thickness are 8.3 Å and 11.3 Å, respectively. The electrodes consist of CNT(8,4) resting on

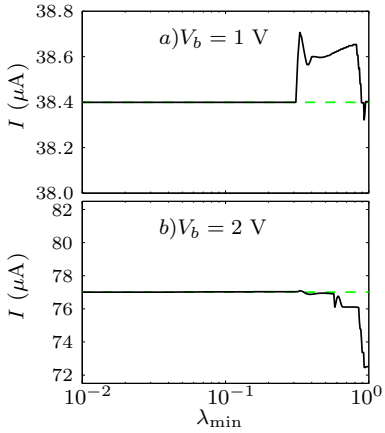


Figure 6.8: Current as a function of the parameter $\lambda_{\min}$ used by the Krylov subspace method for the Au(100)–C7–Au(100) system with applied bias voltages (a) $V_b = 1$ V and (b) $V_b = 2$ V. The correct currents obtained by conventional methods are $I \approx 38.4$ μA and $I \approx 77.0$ μA, respectively, indicated here by the green/dashed lines.

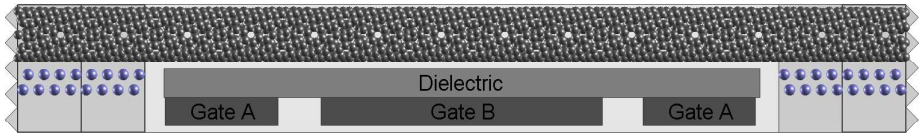


Figure 6.9: (Color online) Schematic illustration of a carbon nanotube (8,4) band-to-band tunneling device. The carbon nanotube is positioned on Li surfaces next to an arrangement of three gates.

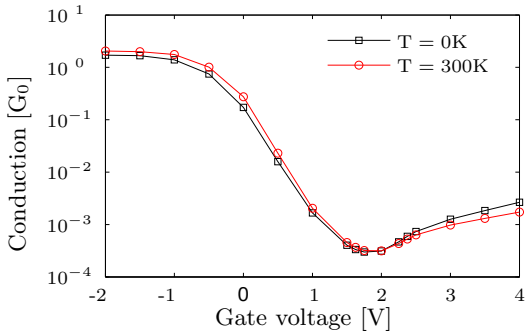


Figure 6.10: (Color online) Conduction as a function of the Gate-A voltage in units of the conductance quantum G_0. In the calculations we use a dielectric constant of 4, $V_{\mathrm{Gate-A}} = -2.0$ V, and vary $V_{\mathrm{Gate-B}}$ from -2.0 V to 4.0 V as indicated.

a thin surfaces of Li, where the lattice constant of the Li layers is stretched to fit the layer thickness of the CNT. The central region of the two-probe system comprises a total of 1440 atoms. An arrangement of rectangular gates are positioned below the carbon nanotube as indicated on the figure. In the plane of the illustration (length × height) the dimensions are as follows: Dielectric 108 Å × 5 Å; Gate-A 108 Å × 5 Å; Gate-B 20 Å × 5 Å. All the regions are centered with respect to the electrodes so that the complete setup has mirror symmetry. In the direction perpendicular to the illustration the configuration is assumed repeated every 19.5 Å as a super-cell.

In the following we show results from a calculation of the transmission spectrum $T(E)$ for $V_{\mathrm{Gate-A}} = -2.0$ V and a dielectric constant of 4. To begin with we calculate the electronic conductance for different Gate-B voltages in the range $[-2\ \mathrm{V}, 4\ \mathrm{V}]$. The results for temperature T = 0 K, in terms of the unit conduction G_0 are displayed with the black curve in Fig. 6.10. It shows an initial conductance for $V_{\mathrm{Gate-B}} = -2.0$ V of the order of one, a subsequent drop by four orders of magnitude around $V_{\mathrm{Gate-B}} = 2.0$ V, and a final increase of one order of magnitude towards $V_{\mathrm{Gate-B}} = 4.0$ V. In addition to the zero temperature conduction which is equal to $T(E_F)$, where E_F is the Fermi energy, we also display the results at room temperature T = 0 K (red curve), which can be obtained from linear response as

$$\mathrm{G} = \int \mathrm{d}E\ T(E) \frac{e^{(E-E_F)/\mathrm{k}_B\mathrm{T}}}{(1+e^{(E-E_F)/\mathrm{k}_B\mathrm{T}})^2} \tag{6.24}$$

The overall trend of the conduction curve is similar for room temperature, and can be explained as band-to-band tunneling which is tuned by the gate potentials.

In order for BTB tunneling to appear in CNFETs, fields along the length of the tube have to be created that are strong enough to shift the conductance or valence bands by at least the gap energy of the CNT. In the case of CNT(8,4) the band gap is ~ 0.8 eV which can be transcended via the three-gate arrangement. More specifically, we present in the left part of Fig. 6.11 the total potential induced by the three gates on the carbon atoms in CNT over the full extension of the device. Along with this, in the right part of Fig. 6.11, we show the corresponding transmission spectrum $T(E)$, for four gate voltages $V_{\mathrm{Gate-B}} = -2.0$ V, 1.0 V, 2.0 V, and 4.0 V, which represent significantly different locations on the conduction curve above.

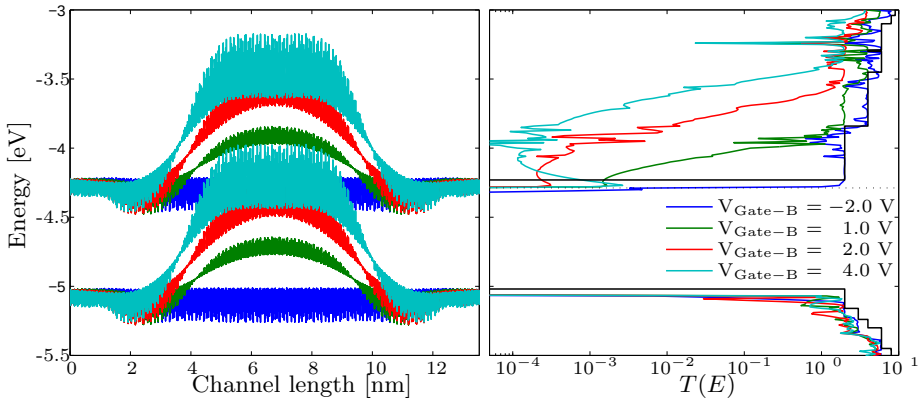


Figure 6.11: (Color online) Fields induced along the length of the device (left panel) and the transmission spectrum (right panel) for the various gate voltages $V_{\mathrm{Gate-B}}$ from -2.0 V to 4.0 V as indicated.

From Fig. 6.11 we can see how the bands are shifted upwards by an increasing amount as the Gate-B voltage in turned up. To begin with, e.g., for $V_{\mathrm{Gate-B}} = 1$ V, this results in lower conduction since the conduction band bends away from the Fermi level, which is indicated by the dotted line. When the gate voltage is at $V_{\mathrm{Gate-B}} = 2$ V the valence band almost reaches the conduction band in which case BTB tunneling becomes possible. By increasing the gate voltage further, more bands become available for BTB tunneling and the effect is visible as a steady increase in the calculated transmission $T(E)$ just above the Fermi level. This behavior is similar to that found in Ref. [89] from experiments in the laboratory with CNFET setups.

6.4.4 CPU runtimes

In this last numerical example we focus on the typical savings in the computational time that can be achieved when computing the self-energy matrices $\mathbf{\Sigma}_L$

Table 6.2: CPU times in seconds for computing the left self-energy matrix $\mathbf{\Sigma}_L$ at twenty different energies E between -2eV and 2eV for selected electrode types and matrix sizes N. The parameter $\lambda_{\min}$ was set to 0.1.

Electrode type	Size	2^n–recursion	DGEEV	Krylov
Li	16	0.1	0.0	0.0
Fe	54	4.2	2.3	0.6
Al(100)	72	4.9	3.3	0.8
Al(100)	128	27.9	17.5	3.6
Au(111)	243	167.2	73.7	11.5
(2, 2)–CNT	64	3.6	2.4	0.7
(4, 4)–CNT	128	26.0	14.4	2.9
(8, 8)–CNT	256	208.8	118.8	17.0
(12, 12)–CNT	384	608.4	373.6	45.6
(16, 16)–CNT	512	1230.0	1403.9	121.5
(20, 20)–CNT	640	1542.3	1125.7	148.0

and $\mathbf{\Sigma}_R$ with the proposed Krylov subspace method. We will compare runtimes directly with conventional schemes usually applied in electron transport calculations.

Table 6.2 presents the profiling results when applying three different methods to calculate the left self-energy matrix $\mathbf{\Sigma}_L$ for common types of electrodes and various matrix sizes N. In every case we consider only the Γ-point and use single-ζ basis sets, except for Au(111) where double-ζ-polarized is used. Since the computational cost might vary significantly with E, the seconds listed is the accumulated time of 20 independent calculations at equidistant energies in the interval $E \in [-2\text{ eV}, 2\text{ eV}]$. In all cases of the Krylov method the parameter $\lambda_{\min}$ was set to 0.1.

From the profiling results in Table 6.2 we see that the computational time of the Krylov subspace method is significantly reduced compared with the presently widely used 2^n–recursive technique. Also the conventional WFM scheme using DGEEV is typically faster than the 2^n–recursive algorithm (the exception for (16, 16)–CNT is related to cache usage[1]). Comparing the timings in the last two columns verifies that the cost to evaluate the self-energy matrices from only the few important Bloch modes of the electrodes, as done in our Krylov subspace method, is in general much lower than required by a direct eigensolver to determine all possible modes.

In order to illustrate the computational complexity of the methods we show

[1]For the armchair (16, 16)–CNT electrode ($N = 512$) the call to DGEEV produces an extremely high number of L2 cache misses, many more than for the bigger (18, 18)–CNT electrode ($N = 576$). This gives the bad times of the DGEEV method for this particular case.

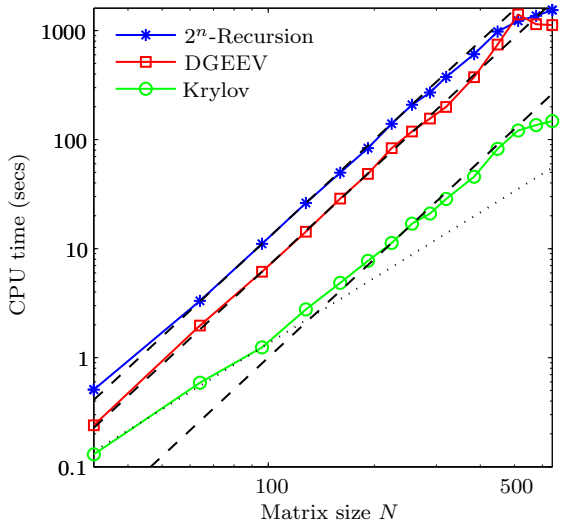


Figure 6.12: CPU times for computing the left self-energy matrix $\mathbf{\Sigma}_L$ plotted as a function of the size N of $\mathbf{\Sigma}_L$ for a range of armchair (n, n)–CNT electrodes, where $n = 1, \ldots, 20$. The dotted and dashed lines indicate $O(N^2)$ and $O(N^3)$ computational complexity, respectively.

the CNT runtimes as a function of the matrix size N in a logarithmic plot in Fig. 6.12. Clearly, all methods have $O(N^3)$ complexity, however, the Krylov subspace method initially follows the typical $O(N^2)$ complexity of the Arnoldi procedure [90] until the cost of the shift-and-invert operations (see Sec. 6.3.3) becomes dominant. For $N > 500$ we observe effects due to more and sometimes less favorable cache usage. Overall, we see that the Krylov subspace method is fastest by an order of magnitude for all but the smallest cases.

It is important to point out that the obtained self-energy matrices $\mathbf{\Sigma}_L$ are in all cases applied in a subsequent transmission calculation of $T(E)$ for the two-probe systems indicated, and the results then checked against those of the conventional methods (the resulting transmissions $T(E)$ are identical for the three methods in all cases of E to at least 3 decimals). Furthermore, the setting of parameter $\lambda_{\min}$ to 0.1 yields self-energy matrices evaluated from all the modes that have phases λ_k satisfying $0.1 < |\lambda_k| < 1 + \epsilon$. This is more than adequate for obtaining correct results for all the systems considered in this section. In practice, the parameter $\lambda_{\min}$ can often be selected > 0.1 without sacrificing any noticeable accuracy in the $T(E)$ calculation, and this would show off the approach as even faster.

Chapter 7

Conclusion and outlook

The subject of this thesis has been the numerical calculation of the electronic properties and in particular the quantum transport properties of devices at the nano-scale. The key method used for such calculations is the Green's function method, which is rooted in the general formalism of the NEGF approach. Alternatively, the method of wave function matching can be used. In order to apply these methods to semiconductor device simulation, however, it is necessary to handle systems comprising millions of atoms, and this will require new efficient algorithms for the most time consuming stages of the calculations.

The primary objective in this thesis has been to develop and implement new algorithms which are faster than existing techniques. We will now briefly summarize the main features of the resulting algorithms and present an outlook on future work related to their application.

- **Block tridiagonal matrix inverse** (algorithm viii): The block tridiagonal part of the Green's function matrix for the central region of the nano-scale system is obtained as a matrix inverse operation in $O(N)$ computational complexity. This algorithm is based on two independent Block Gaussian elimination sweeps and can be run in parallel on two CPUs. We expect this algorithm to be a key ingredient in the further parallelization of the matrix inverse in combination with a block cyclic reduction technique. This is ongoing research at the moment.

- **Efficient wave function matching method** (algorithm x): A modified WFM approach is developed that allows for the exclusion of the majority of the evanescent modes of the bulk electrodes in all parts of a transmission spectrum calculation. The approach makes it feasible to apply iterative techniques to efficiently determine the relatively few bulk modes of interest, which allows for a significant reduction of the computational expense in practice. We believe this scheme has great potential for studying electronic transport in large-scale atomic two-probe systems, such as large carbon nanotubes or nano-wire configurations. In addition, the real power of the approach - something which we have not addressed

in this work - is its application to obtain the density matrix during the self-consistent procedure in non-equilibrium situations. Our preliminary testing shows a speed-up of more than 20 in the most costly part of the calculation. We expect to study this further in the near future.

- **Krylov subspace method** (ALGORITHM XII): A procedure based on the Arnoldi method is employed to obtain solutions of quadratic eigenvalue problems. One complex and two real shift-and-invert transformations are adopted to select interior eigenpairs with complex eigenvalues on or close to the unit circle that correspond to the propagating and evanescent modes required in our efficient WFM method. The algorithm is robust and much faster than conventional LAPACK routines employing direct eigensolvers. The execution of the algorithm can be parallelized over the independent shift-and-invert stages (four in the general complex case). Other iterative methods which might avoid the explicit shift-and-invert operations, such as the Jacobi-Davidson techniques, have potential to outperform the proposed method, in particular on a parallel platform. We leave this to future work.

Appendix A

Green's functions

In this appendix, we briefly summarize some of the basic properties of Green's functions. First the fundamental mathematical properties are reviewed which leads to the appropriate physical interpretation that is used in electronic transport theory. Then the Green's function for the simple, constant potential 1D wire example is derived and this is subsequently used to generalize the ideas and find the Green's function solution for the 3D electrodes used in our two-probe setups.

A.1 Mathematical properties

Mathematically, a Green's function is a function that can be used to solve inhomogeneous differential equations subject to boundary conditions. For example, consider the general form of a linear differential equation, given by

$$\hat{L}(x)u(x) = f(x), \tag{A.1}$$

where $\hat{L}(x)$ is a linear differential operator, $u(x)$ is an unknown response function, and $f(x)$ is a known non-homogeneous source term. We can write the solution as

$$u(x) = \hat{L}^{-1}(x)f(x) \equiv \int G(x,x')f(x')dx', \tag{A.2}$$

where $\hat{L}^{-1}$ is the inverse of the differential operator $\hat{L}$ and the last definition is reasonable, since we must expect the inverse of a differential operator to be an integral operator. The two-point kernel $G(x,x')$ within the integral is then the *Green's function* associated with the differential operator $\hat{L}$.

The Green's function defined in Eq. (A.2) can be associated with the Dirac delta function $\delta(x)$ that has the well-known property

$$\int \delta(x-x')f(x')dx' = f(x), \tag{A.3}$$

which leads directly to the important relation

$$\hat{L}(x)G(x,x') = \delta(x-x'), \tag{A.4}$$

when comparing with the result from inserting Eq. (A.2) into Eq. (A.1) and moving the linear operator $\hat{L}$ inside the integral.

We note that the inverse of a differential operator, and thus $G(x,x')$, is not uniquely specified till we specify the boundary conditions. In addition, the Green's function can be proved to have the following mathematical properties:

- $G(x,x')$ satisfies the homogeneous differential equation $\hat{L}(x)G(x,x') = 0$ at all points other than $x = x'$.
- $G(x,x')$ is continues at $x = x'$.
- $\frac{dG(x,x')}{dx}$ is discontinues at $x = x'$.
- $G(x,x')$ is symmetrical with respect to x and x'.

A.2 Infinite 1D wire

The movement of the electrons in the two-probe systems discussed in this thesis is described by the single-particle Hamiltonian operator, given by

$$\hat{H} = \Big[-\frac{\hbar^2}{2m_e}\nabla^2 + V(\mathbf{r})\Big]. \tag{A.5}$$

Since this is a differential operator one can solve the corresponding Schrödinger equation $[\hat{H} - E]\Psi(\mathbf{r}) = 0$, by finding the associated Green's function, that is

$$G(\mathbf{r},\mathbf{r}') = [\hat{H} - E]^{-1}, \quad [\hat{H} - E]G(\mathbf{r},\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}'), \tag{A.6}$$

To illustrate, let's consider a simple 1D case of the Hamiltonian in Eq. (A.5), where the potential is just a constant $V(\mathbf{r}) \to V_0$ and find the Green's function for

$$\Big[-\frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial x^2} + V_0 - E\Big]G(x,x') = \delta(x-x'). \tag{A.7}$$

Clearly, the homogeneous version of this equation resembles a one-dimensional Schrödinger equation and will have solutions of wave form $\sim e^{\hat{\imath}kx}$. Since $G(x,x')$ should satisfy the homogeneous equation at all x except $x = x'$, we can write

$$G(x,x') = \begin{cases} A^+ e^{\hat{\imath}k(x-x')}, & x > x' \\ A^- e^{-\hat{\imath}k(x-x')}, & x < x' \end{cases} \tag{A.8}$$

where A^+ and A^- are constants and $k = [2m(V_0 - E)]^{1/2}/\hbar$. No matter what A^+ and A^- might be, this solution satisfies Eq. (A.7) at all points other than

$x = x'$. Taking into account the two mathematical properties, that $G(x, x')$ is continues and $\frac{dG(x,x')}{dx}$ is discontinues (in this case by $2m_e/\hbar^2$) at $x = x'$, we can determine the amplitudes $A^+ = A^- = -im_e/k\hbar^2$, which gives the solution

$$G(x, x') = -\frac{im_e}{k\hbar^2} e^{\hat{\imath}k|x-x'|} \tag{A.9}$$

valid for all x and x'. From a mathematically point of view, we immediately note that a solution with $k \to -k$ satisfies Eq. (A.7) just as well.

A.3 Physical interpretation

From a physical point of view, we interpret the Green's function $G(x, x')$ defined in Eq. (A.7) as the wave function at x resulting from a unit excitation applied at x'. As the above example shows, we can expect such an excitation in a 1D wire to give rise to two waves traveling outwards from the point of excitation with equal amplitudes A^+ and A^-. Mathematically, however, there exists another solution, which corresponds to incoming waves that disappear at the point of excitation. The two solutions are referred to as the advanced and retarded Green's functions, respectively, and satisfy the same differential equation but with different boundary conditions. The names are linked to the Green's function representation in the time domain as obtained from a Fourier transformation $G(t) = \int \frac{dE}{2\pi\hbar} e^{+iEt/\hbar} G(E)$. The retarded solution is zero for $t < 0$ (causal) and interpreted as the response of an excitation at $t = 0$, while the advanced is zero for $t > 0$ (non-causal), with no direct physical meaning [4].

We can incorporate the boundary conditions into the the Green's function definition itself by adding a small imaginary part to the energy, i.e. $E \to E + \hat{\imath}\eta$, where η is an infinitesimal. In the 1D wire example, this yields the equation

$$\Big[-\frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial x^2} + V_0 - (E + \hat{\imath}\eta) \Big] G^r(x, x') = \delta(x - x'). \tag{A.10}$$

for $\eta > 0$ where the superscript "r" on $G^r(x, x')$ indicates that this is only valid for the retarded solution. It is easy to see this from physical arguments since the wave number k^r for the Green's function solution to Eq. (A.10) can be written

$$k^r = \frac{\sqrt{2m(V_0 - (E + \hat{\imath}\eta))}}{\hbar} = \frac{\sqrt{2m(V_0 - E)}}{\hbar}\sqrt{1 + \frac{\hat{\imath}\eta}{V_0 - E}} \approx k(1 + i\tilde{\eta}) \tag{A.11}$$

where $\tilde{\eta} = \eta/2(V_0 - E)$, which indicates, that the wave number has gained a positive imaginary component compared to the original wave number k. This imaginary part makes the advanced solution, given by $k \to -k^r$ in Eq. (A.9), grow indefinitely as we move away from the point of excitation and it is therefore not physically acceptable. The retarded solution, on the other hand, is well bounded as it decreases infinitesimally. If we had subtracted $\hat{\imath}\eta$ from the energy instead, i.e. $E \to E - \hat{\imath}\eta$, the opposite would have been the case, and then the advanced Green's function $G^a(x, x')$ would be the only correct solution.

A.4 Ideal layered 3D electrodes

The 3D electrodes modeled in this thesis have a finite size in two dimensions and can be divided into principal cells in the third and infinite (transmission) dimension, as described in detail, e.g., in Sec. 3.2.2. Furthermore, our formalism assumes that the wave function in the ith cell can be written $\Psi_i = \sum_{m=1}^{M_i} c_{i,m}\phi_{i,m}$, where $c_{i,m}$ are expansion coefficients and $\phi_{i,m}$ is the $m = 1, \ldots, M_i$ atomic orbitals local to cell i. The Schrödinger equation for this system becomes block tridiagonal as shown in Sec. 3.1.4. Here we will consider the ideal electrode where all the diagonal and coupling blocks of $\bar{\mathbf{H}}$ are the same, i.e. $\bar{\mathbf{H}}_i \equiv \bar{\mathbf{H}}_L$ and $\bar{\mathbf{H}}_{i,i+1} \equiv \bar{\mathbf{H}}_{L,L}$, and derive the corresponding Green's function solution.

From comparison with the infinite 1D wire example, it is reasonable to expect that the Green's function is given by an infinite matrix just like the Hamiltonian and that it satisfies the matrix equation

$$\begin{pmatrix} \ddots & \ddots & & \\ \ddots & \bar{\mathbf{H}}_L & \bar{\mathbf{H}}_{L,L} & \\ & \bar{\mathbf{H}}_{L,L}^\dagger & \bar{\mathbf{H}}_L & \ddots \\ & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} & \vdots & \vdots & \vdots & \\ \cdots & \mathbf{G}_{i-1,i-1} & \mathbf{G}_{i-1,i} & \mathbf{G}_{i-1,i+1} & \cdots \\ \cdots & \mathbf{G}_{i,i-1} & \mathbf{G}_{i,i} & \mathbf{G}_{i,i+1} & \cdots \\ \cdots & \mathbf{G}_{i+1,i-1} & \mathbf{G}_{i+1,i} & \mathbf{G}_{i+1,i+1} & \cdots \\ & \vdots & \vdots & \vdots & \end{pmatrix} = \mathbf{I}, \tag{A.12}$$

where $\mathbf{I}$ is the infinite identity matrix. The physical interpretation of this equation and $\mathbf{G}$ is the same as before. However, the Green's function is now a compound object that includes all wave functions resulting from all unit excitations applied not at a single point, but over the spatial extend of an atomic orbital. Each diagonal entry in the infinite matrix $\mathbf{I}$ corresponds to an excitation at a localized orbital of the system. This means, that the block $\mathbf{G}_{i,j}$ gives the wave functions in cell i resulting from the possible excitations in cell j.

Some important properties of the Green's function matrix follow directly from the definition Eq. (A.12). First we note, that since the infinite Hamiltonian is a toeplitz block matrix, then so is it's inverse $\mathbf{G}$, if it exists. This means, that all blocks $\mathbf{G}_{i,j}$ are the same for the same distance $i - j$ from the diagonal and that the full solution is fixed by a single column of $\mathbf{G}$. So we can limit our solution to column j of $\mathbf{G}$ for which the matrix equation Eq. (A.12) simplifies to

$$\bar{\mathbf{H}}_{L,L}^\dagger \mathbf{G}_{i-1,j} + \bar{\mathbf{H}}_L \mathbf{G}_{i,j} + \bar{\mathbf{H}}_{L,L} \mathbf{G}_{i+1,j} = \delta_{i,j}\mathbf{I}. \tag{A.13}$$

for all $i = -\infty, \ldots, \infty$. Clearly, this simplification also follows from physical symmetry arguments.

In the same manner as for the 1D wire example we can write down a general solution for the retarded Green's function block $\mathbf{G}_{i,j}$. Since an excitation should

create two waves traveling outward from the cell j of excitation, it follows that

$$\mathbf{G}^r_{i,j} = \begin{cases} \mathbf{C}^+_{i,j}\mathbf{A}^+, & i > j \\ \mathbf{C}^-_{i,j}\mathbf{A}^-, & i < j \end{cases}, \tag{A.14}$$

where $\mathbf{A}^\pm$ are diagonal matrices of constants and $\mathbf{C}^\pm_{i,j}$ are matrices having vectors as columns that hold the expansion coefficients for the right-going (+) and left-going (−) waves of the cell i. We write the $\mathbf{C}^\pm_{i,j}$ matrices as

$$\mathbf{C}^+_{i,j} = \left(\mathbf{c}^{1+}_{i,j}, \ldots, \mathbf{c}^{M_j+}_{i,j}\right), \quad \mathbf{C}^-_{i,j} = \left(\mathbf{c}^{1-}_{i,j}, \ldots, \mathbf{c}^{M_j-}_{i,j}\right), \tag{A.15}$$

where the m^{th} column of these matrices ($\mathbf{c}^{m\pm}_{i,j}$) will hold the coefficients of the wave function resulting from a unit excitation of the orbital $\phi_{j,m}$ in cell j.

Since the system is periodic in the transmission direction the wave functions of the principal cells only differ by a phase factor according to Bloch's theorem, i.e. $\Psi_i = e^{\hat{\imath}kd}\Psi_{i-1}$. This allows us to relate adjacent $\mathbf{C}^\pm_{i,j}$ matrices by

$$\mathbf{C}^\pm_{i,j} = \mathbf{C}^\pm_{i-1,j}\mathbf{\Lambda}^\pm \tag{A.16}$$

where $\mathbf{\Lambda}^\pm$ are diagonal matrices having the phase factors $\lambda^\pm_m \equiv e^{\hat{\imath}k^\pm_m d}$ down the diagonal. Here $k^\pm_m$ is the wave number corresponding to the wave with $\mathbf{c}^{m\pm}_{i-1,j}$ as coefficients (the m^{th} mode) and d is the distance between the cells.

We now note, that if the diagonal block $\mathbf{G}^r_{i,i}$ is known, then the entire solution $\mathbf{G}^r$ in Eq. (A.14) for all i and j is available by using the toeplitz property and the Bloch relation in Eq. (A.16). Fortunately, an expression for the diagonal block is easily obtained by setting $j = i$ in Eq. (A.13) and then inserting

$$\begin{aligned} \mathbf{G}^r_{i+1,i} &= \mathbf{C}^+_{i+1,i}\mathbf{A}^+ = \mathbf{C}^+_{i,i}\mathbf{\Lambda}^+\mathbf{A}^+ = \mathbf{C}^+_{i,i}\mathbf{\Lambda}^+\tilde{\mathbf{C}}^+_{i,i}\mathbf{C}^+_{i,i}\mathbf{A}^+ = \mathbf{B}^+\mathbf{G}^r_{i,i}, \quad \text{and} \\ \mathbf{G}^r_{i-1,i} &= \mathbf{C}^-_{i-1,i}\mathbf{A}^- = \mathbf{C}^-_{i,i}(\mathbf{\Lambda}^-)^{-1}\mathbf{A}^- = \mathbf{C}^-_{i,i}(\mathbf{\Lambda}^-)^{-1}\tilde{\mathbf{C}}^-_{i,i}\mathbf{C}^-_{i,i}\mathbf{A}^- = (\mathbf{B}^-)^{-1}\mathbf{G}^r_{i,i}, \end{aligned} \tag{A.17}$$

where the matrices $\tilde{\mathbf{C}}^\pm_{i,i}$ are the inverses of $\mathbf{C}^\pm_{i,i}$, making $\tilde{\mathbf{C}}^\pm_{i,i}\mathbf{C}^\pm_{i,i} = \mathbf{C}^\pm_{i,i}\tilde{\mathbf{C}}^\pm_{i,i} = \mathbf{I}$, and in the last step we have introduced the so-called Bloch matrices defined by

$$\mathbf{B}^\pm \equiv \mathbf{C}^\pm\mathbf{\Lambda}^\pm\tilde{\mathbf{C}}^\pm, \tag{A.18}$$

leaving out the implied subscripts i, i. A key assumption in chapters 5 and 6 of this thesis is that we can obtain approximate solutions for $\mathbf{G}^r$ by omitting the less important columns in the $\mathbf{C}^\pm$ matrices. In such a case, $\tilde{\mathbf{C}}^\pm$ denotes the (Moore-Penrose) pseudo-inverses of $\mathbf{C}^\pm$, and $\mathbf{C}^\pm\tilde{\mathbf{C}}^\pm \neq \mathbf{I}$.

All blocks on the diagonal of the infinite retarded Green's function matrix are then given by

$$\mathbf{G}^r_{i,i} = \left[\bar{\mathbf{H}}^\dagger_{L,L}(\mathbf{B}^-)^{-1} + \bar{\mathbf{H}}_L + \bar{\mathbf{H}}_{L,L}\mathbf{B}^+\right]^{-1} = \left[\bar{\mathbf{H}}^\dagger_{L,L}\left[(\mathbf{B}^-)^{-1} - (\mathbf{B}^+)^{-1}\right]\right]^{-1} \tag{A.19}$$

where it was used that $\bar{\mathbf{H}}_{L,L}^{\dagger}(\mathbf{B}^{\pm})^{-1}+\bar{\mathbf{H}}_{L}+\bar{\mathbf{H}}_{L,L}\mathbf{B}^{\pm}=0$, which follows from the Schrödinger equation for the system. Equivalently, $\mathbf{G}_{i,i}^{r}=\left[\bar{\mathbf{H}}_{L,L}(\mathbf{B}^{+}-\mathbf{B}^{-})\right]^{-1}$ is then also satisfied. The recursion relations in Eq. (A.17) now yield all other blocks of the Green's function matrix.

Appendix B

The NEGF formalism for coherent transport

The non-equilibrium Green's function (NEGF) formalism is a conceptual and computational framework for treating quantum transport in nano-scale devices under finite bias. In its most complete form, it goes beyond the Landauer approach for coherent conduction, to include inelastic scattering and strong correlation effects at an atomistic level [4, 38, 44]. The aim of this appendix is to obtain the main formulas of the NEGF theory in the coherent limit, which is considered in the current thesis. We will follow the description and notation of Datta from his famous textbooks [4, 91].

B.1 Electron reservoir probes

In the NEGF approach the electrodes (probes) constitute reservoirs of electrons held a certain electrochemical potentials. Consider for example a semi-infinite (left) electrode as the one depicted in the left part of Fig. 3.4. Again, assume a separation into principal layers of the electrodes bulk material limiting the interaction between layers to next neighbors. Electrons in this isolated electrode will satisfy the Schrödinger equation, in matrix form, that is (cf. Eq. (3.10))

$$[E\mathbf{S}_L^\infty - \mathbf{H}_L^\infty]\mathbf{c}_L = 0, \tag{B.1}$$

where the semi-infinite Hamiltonian of the electrode is

$$\mathbf{H}_L^\infty = \begin{pmatrix} \ddots & \ddots & & \\ \ddots & \mathbf{H}_L & \mathbf{H}_{L,L} & \\ & \mathbf{H}_{L,L}^\dagger & \mathbf{H}_L & \mathbf{H}_{L,L} \\ & & \mathbf{H}_{L,L}^\dagger & \mathbf{H}_L \end{pmatrix}, \tag{B.2}$$

and $\mathbf{S}_L^\infty$ is the corresponding overlap matrix. The superscript ∞ indicates that there is an infinite number of principal layers in the electrode Hamiltonian. As

explained in detail by Datta [4], these equations cannot describe a reservoir at a constant electrochemical potential since this requires a form that allows continues extraction and re-injection of electrons from external sources. We therefore modify equation Eq. (B.1) by adding a constant perturbation and write

$$[(E+\hat{\imath}\eta)\mathbf{S}_L^\infty - \mathbf{H}_L^\infty]\mathbf{c}_L = \mathbf{Q}_L, \tag{B.3}$$

where $\eta = 0^+$ is a small positive infinitesimal number that effectively introduces extraction of electrons and the term $\mathbf{Q}_L$ on the right represents an external source re-injecting electrons. One should note, that E is no longer an eigenvalue of $(\mathbf{S}_L^\infty)^{-1}\mathbf{H}_L^\infty$ at which the electron wave functions $\mathbf{c}_L$ exist as eigenfunctions. Now the $\mathbf{c}_L$'s will be non-zero at all energies E, having peaks (whose sharpness depends on the value of η) around the eigenvalues of $(\mathbf{S}_L^\infty)^{-1}\mathbf{H}_L^\infty$.

B.2 One-probe setup

Before turning to the interesting two-probe system it is convenient to investigate the simpler case of a central system connected to only one electrode. This situation is illustrated as the L-C part of Fig. 3.4, and using Eq. (B.3) we can write the Schrödinger equation for the composite system in two-block form, given by

$$\begin{pmatrix} (E+\hat{\imath}\eta)\mathbf{S}_L^\infty - \mathbf{H}_L^\infty & \bar{\mathbf{H}}_{L,C}^\infty \\ \bar{\mathbf{H}}_{L,C}^{\infty\dagger} & \bar{\mathbf{H}}_C \end{pmatrix} \begin{pmatrix} \mathbf{c}_L + \mathbf{c}_r \\ \mathbf{c}_C \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_L \\ \mathbf{0} \end{pmatrix}, \tag{B.4}$$

where $\mathbf{H}_C$ is the Hamiltonian for the central region, $\mathbf{H}_{L,C}^\infty$ is the coupling matrix between the electrode and the central region, and the notation $\bar{\mathbf{H}}_C \equiv E\mathbf{S}_C - \mathbf{H}_C$ and $\bar{\mathbf{H}}_{L,C}^\infty \equiv E\mathbf{S}_{L,C}^\infty - \mathbf{H}_{L,C}^\infty$ was adopted. The $\mathbf{c}_C$ vector holds the expansion coefficients of the wave function that is created inside the central region as a consequence of it's contact with the electrode [1]. In turn this excites scattered waves in the central region that are reflected back into the electrode, here represented by the coefficients $\mathbf{c}_C^{\text{ref}}$.

Assuming that the source term $\mathbf{Q}_L$ in the upper line of the block equation Eq. (B.4) is constant and the same as in Eq. (B.3), we can easily eliminate it from this line. Then it is possible to express $\mathbf{c}_C^{\text{ref}}$ in terms of $\mathbf{c}_C$, that is

$$[(E+\hat{\imath}\eta)\mathbf{S}_L^\infty - \mathbf{H}_L^\infty]\mathbf{c}_C^{\text{ref}} + \bar{\mathbf{H}}_{L,C}^\infty \mathbf{c}_C = 0 \quad \Leftrightarrow \quad \mathbf{c}_C^{\text{ref}} = -\mathbf{G}_L \bar{\mathbf{H}}_{L,C}^\infty \mathbf{c}, \tag{B.5}$$

where the semi-infinite Green's function of the left electrode

$$\mathbf{G}_L \equiv [(E+\hat{\imath}\eta)\mathbf{S}_L^\infty - \mathbf{H}_L^\infty]^{-1} \tag{B.6}$$

has been defined. Substituting the expression for $\mathbf{c}_C^{\text{ref}}$ in Eq. (B.5) into the lower line of the block equation Eq. (B.3) gives

$$\bar{\mathbf{H}}_{L,C}^{\infty\dagger}\mathbf{c}_L + [\bar{\mathbf{H}}_C - \bar{\mathbf{H}}_{L,C}^{\infty\dagger}\mathbf{G}_L\bar{\mathbf{H}}_{L,C}^\infty]\mathbf{c}_C = 0, \tag{B.7}$$

[1] This is called the "spilling over" of the electrode wave function in the Datta terminology!

which can be rearranged and written as $\mathbf{c}_C = \mathbf{GQ}$, where

$$\mathbf{G} \equiv [\bar{\mathbf{H}}_C - \mathbf{\Sigma}_L]^{-1}, \quad \mathbf{Q} \equiv -\bar{\mathbf{H}}^{\infty\dagger}_{L,C}\mathbf{c}_L, \tag{B.8}$$

and the new quantity

$$\mathbf{\Sigma}_L \equiv \bar{\mathbf{H}}^{\infty\dagger}_{L,C}\mathbf{G}_L\bar{\mathbf{H}}^{\infty}_{L,C}, \tag{B.9}$$

is called the self-energy. As expected, $\mathbf{c}_C = \mathbf{GQ}$ has the form of a Schrödinger equation describing an open system with a source term $\mathbf{Q}$ arising from the reservoir electrode. A derivation for a right electrode yields similar formulas with $\mathbf{Q} = -\bar{\mathbf{H}}^{\infty}_{R,C}\mathbf{c}^{\text{ref}}_C$ and a self-energy, given by $\mathbf{\Sigma}_R \equiv \bar{\mathbf{H}}^{\infty}_{R,C}\mathbf{G}_R\bar{\mathbf{H}}^{\infty\dagger}_{R,C}$.

It is clear, that the size of the coupling matrix $\bar{\mathbf{H}}^{\infty}_{L,C}$ between the left electrode and the central region will be semi-infinite just like the electrode Hamiltonian matrix. This seems to prevent the calculation of the self-energy from eq. Eq. (B.9), however, in our tight-binding setup (see fig. 4), the elements of $\bar{\mathbf{H}}^{\infty}_{L,C}$ are all zero except in the lower left corner where it is equal to $\bar{\mathbf{H}}_{L,L}$ of size $m_L \times m_L$, where $\bar{\mathbf{H}}_{L,L} \equiv (E + \hat{\imath}\eta)\mathbf{S}_{L,L} - \mathbf{H}_{L,L}$. This means we can calculate the non-zero elements of the self-energy matrix from the finite matrix equation

$$[\mathbf{\Sigma}_L]_{m_L \times m_L} = \bar{\mathbf{H}}^{\dagger}_{L,L}\mathbf{g}_L\bar{\mathbf{H}}_{L,L}, \tag{B.10}$$

where $\mathbf{g}_L$ is the so-called surface Green's function of the electrode corresponding to the lower right $m_L \times m_L$ submatrix of $\mathbf{g}_L$.

Notice that the self-energy $\mathbf{\Sigma}_L$ is not hermitian because of $\hat{\imath}\eta$ and as a result the effective Hamiltonian of the composite system $\bar{\mathbf{H}}_C - \mathbf{\Sigma}_L$ is not hermitian either, giving complex eigenvalues. It can be shown, that the real part of the self-energy is responsible for shifting the energy levels of the central Hamiltonian, while the imaginary part can be related to the lifetime of the levels [4]. As the inverse lifetime is proportional to the broadening of a level, the anti-hermitian component of the self-energy, given by

$$\mathbf{\Gamma}_L = \hat{\imath}(\mathbf{\Sigma}_L - \mathbf{\Sigma}^{\dagger}_L), \tag{B.11}$$

is called the broadening matrix and is to be used in the transmission expression in Eq. (B.42) for the left electrode derived below.

B.3 Two-probe setup

We now consider the full two-probe system as shown in figure 5. Looking at the results from the previous section it is reasonable to expect that the new right electrode will influence the equations in a similar manner as the left electrode did. A complete derivation of the two-probe expressions can be found in Datta's book and confirms this anticipation [4]. We can thus immediately write down the expression corresponding to Eq. (B.8) for the resulting two-probe Schrödinger equation by redefining the Green's function $\mathbf{G}$ and the source term $\mathbf{Q}$ as

$$\mathbf{G} = [\bar{\mathbf{H}}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R]^{-1}, \quad \mathbf{Q} = -\bar{\mathbf{H}}^{\infty\dagger}_{L,C}\mathbf{c}_L - \bar{\mathbf{H}}^{\infty}_{R,C}\mathbf{c}^{\text{ref}}_C, \tag{B.12}$$

where $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ are the self-energies for the left and right electrodes, i.e.

$$\mathbf{\Sigma}_L \equiv \bar{\mathbf{H}}_{L,C}^{\infty\dagger}\mathbf{G}_L\bar{\mathbf{H}}_{L,C}^{\infty}, \quad \mathbf{\Sigma}_R \equiv \bar{\mathbf{H}}_{R,C}^{\infty}\mathbf{G}_R\bar{\mathbf{H}}_{R,C}^{\infty\dagger}. \tag{B.13}$$

As discussed for the one-probe setup, the self-energy matrices have only $m_L \times m_L$ and $m_R \times m_R$ non-zero elements, respectively, that in practice are determined by

$$[\mathbf{\Sigma}_L]_{m_L\times m_L} = \bar{\mathbf{H}}_{L,L}^{\dagger}\mathbf{g}_L\bar{\mathbf{H}}_{L,L}, \quad [\mathbf{\Sigma}_R]_{m_R\times m_R} = \bar{\mathbf{H}}_{R,R}\mathbf{g}_R\bar{\mathbf{H}}_{R,R}^{\dagger}, \tag{B.14}$$

where $\mathbf{g}_L$ and $\mathbf{g}_R$ are the surface Green's functions corresponding to the lower right $m_L \times m_L$ submatrix of $\mathbf{g}_L$ and the upper left $m_r \times m_r$ submatrix of $\mathbf{g}_R$, respectively.

In conclusion, we end up with a finite expression for the C region Green's function, which can be written

$$\mathbf{G}_C = \begin{pmatrix} \bar{\mathbf{H}}_1 - \mathbf{\Sigma}_L & \bar{\mathbf{H}}_{1,2} & & \\ \bar{\mathbf{H}}_{1,2}^{\dagger} & \ddots & \ddots & \\ & \ddots & \ddots & \bar{\mathbf{H}}_{\mathrm{n}-1,\mathrm{n}} \\ & & \bar{\mathbf{H}}_{\mathrm{n}-1,\mathrm{n}}^{\dagger} & \bar{\mathbf{H}}_{\mathrm{n}} - \mathbf{\Sigma}_R \end{pmatrix}^{-1}, \tag{B.15}$$

where $\bar{\mathbf{H}}_i$ and $\bar{\mathbf{H}}_{i,i-1}$, $i = 1, \ldots, \mathrm{n}$, are the diagonal and off-diagonal Hamiltonian blocks describing the individual principal layers of C. The self-energy matrices exist only in the corner blocks. This is exactly the same expression as obtained in Eq. (3.23) from simple block Gaussian elimination arguments.

B.4 Spectral function

Another useful quantity in the NEGF formalism is the matrix corresponding to the spectral function, defined as

$$\mathbf{A} = 2\pi \sum_{\mathrm{k}} \mathbf{c}_{\mathrm{k}}\delta(E - \epsilon_{\mathrm{k}})\mathbf{c}_{\mathrm{k}}^{\dagger}, \tag{B.16}$$

where the sum is over wave numbers k that designate the eigenvectors $\mathbf{c}_{\mathrm{k}}$ and eigenvalues ϵ_{k} of the Schrödinger eigenvalue problem for the central region

$$\mathbf{H}'_C\mathbf{c} = \epsilon\mathbf{S}_C\mathbf{c}, \tag{B.17}$$

where $\mathbf{H}'_C = \mathbf{H}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R$ is equal to the right-hand-side matrix of Eq. (B.15) without the inversion operation. Subsequently, the trace of the spectral matrix gives

$$\mathbf{D} = \frac{1}{2\pi}\mathrm{Tr}\{\mathbf{A}\}, \tag{B.18}$$

as the LDOS. Also, notice that all solutions to Eq. (B.17) are available from $\mathbf{A}$, since the vector $\mathbf{c} = \mathbf{A}\mathbf{b}$ is a solution for any choice of $\mathbf{b}$.

Suppose that the spectral function is represented in the eigenbasis of $\mathbf{H}'_C$, i.e., in the set of solutions $\{\mathbf{c}_\mathrm{k}\}$. In this basis, Eq. (B.18) can be written as

$$\mathbf{A} = 2\pi\delta(E\mathbf{S}_C - \mathbf{H}'_C), \tag{B.19}$$

since $\mathbf{S}_C$ is then the identity matrix and both $\mathbf{H}'_C$ and $\mathbf{A}$ are diagonal matrices with elements ϵ_k and $\delta(E - \epsilon_\mathrm{k})$ on the diagonal, respectively (i.e., the notation $\delta(E\mathbf{S}_C - \bar{\mathbf{H}}'_C)$ means that δ works on each diagonal element separately). We can now use the following identity for the Dirac delta function,

$$\delta(x) = \frac{1}{\pi}\lim_{\eta\to 0}\frac{\eta}{x^2+\eta^2} = \frac{\hat{\imath}}{2\pi}\lim_{\eta\to 0}\left(\frac{1}{x+\hat{\imath}\eta} - \frac{1}{x-\hat{\imath}\eta}\right), \tag{B.20}$$

to write

$$\mathbf{A} = \hat{\imath}\left([(E+\hat{\imath}\eta)\mathbf{S}_C - \mathbf{H}'_C)]^{-1} - [(E-\hat{\imath}\eta)\mathbf{S}_C - \mathbf{H}'_C)]^{-1}\right), \tag{B.21}$$

where $\eta = 0^+$ is a positive infinitesimal number. It can be shown that the above matrix equation is valid not only in the eigenbasis representation but in any representation [4]. Consequently, by comparison with Eq. (B.12), we arrive at an elegant expression for the spectral function in terms of Green's functions:

$$\mathbf{A} = \hat{\imath}\left[\mathbf{G}^r_C - \mathbf{G}^a_C\right], \tag{B.22}$$

where the retarded and advanced Green's functions of the C region are defined

$$\mathbf{G}^r_C = [(E+\hat{\imath}\eta)\mathbf{S}_C - \mathbf{H}'_C)]^{-1}, \tag{B.23}$$

and

$$\mathbf{G}^a_C = [(E-\hat{\imath}\eta)\mathbf{S}_C - \mathbf{H}'_C)]^{-1}, \tag{B.24}$$

respectively. These matrices will have the same overall structure as given in Eq. (B.15) but with added imaginary parts $\pm\hat{\imath}\eta$ to the energy. It is implicitly assumed in the above derivation that the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$, which enters in Eqs. (B.23) and (B.24) through $\mathbf{H}'_C$, are defined with the same imaginary part $\pm\hat{\imath}\eta$ as is explicit in the expressions for $\mathbf{G}^r_C$ and $\mathbf{G}^a_C$. We also note that $\mathbf{G}^a_C = (\mathbf{G}^r_C)^\dagger$ for real E in this case, since $\mathbf{\Sigma}^a = (\mathbf{\Sigma}^r)^\dagger$ and $\mathbf{S}_C$ and $\mathbf{H}_C$ are symmetric [4].

B.5 Electron density

Consider now the wave function of the central region in the NEGF formalism,

$$\mathbf{c}_C = \mathbf{G}_C\mathbf{Q} = -\mathbf{G}_C(\bar{\mathbf{H}}^{\infty\dagger}_{L,C}\mathbf{c}_L + \bar{\mathbf{H}}^{\infty}_{R,C}\mathbf{c}^{\mathrm{ref}}_C) = \mathbf{c}^+_C + \mathbf{c}^-_C, \tag{B.25}$$

where $\mathbf{Q}$ is the source term from Eq. (B.12) in the two-probe case. The wave function is separated into independent parts originating from the left electrode (+) and a the right electrode (−), respectively. Assuming that the electrons are from corresponding electron reservoirs (like in the Landauer picture of Sec. 3.1.2), the electron density matrix can be written as

$$\mathbf{n} = \mathbf{n}^+ + \mathbf{n}^- = 2\int_{-\infty}^{\infty} \frac{1}{2\pi}\left(\mathbf{A}^+ f(E-\mu_L) + \mathbf{A}^- f(E-\mu_R)\right)\,\mathrm{d}E \tag{B.26}$$

where the factor 2 is from spin degeneracy, f is the Fermi function, and

$$\mathbf{A}^{\pm} = 2\pi \sum_{\mathrm{k}} \mathbf{c}^{\pm}_{C,\mathrm{k}} \delta(E - E^{\pm}_{\mathrm{k}}) \mathbf{c}^{\pm\dagger}_{C,\mathrm{k}} \tag{B.27}$$

are the spectral matrices corresponding the states originating from left (+) and right (−). Furthermore, inserting explicitly the expressions for $\mathbf{c}^+_C$ and $\mathbf{c}^-_C$ in Eq. (B.25) into Eq. (B.27), yields

$$\begin{aligned}
\mathbf{A}^+ &= 2\pi \sum_{\mathrm{k}} \mathbf{G}_C \bar{\mathbf{H}}^{\infty\dagger}_{L,C} \mathbf{c}_{L,\mathrm{k}} \delta(E - E_{\mathrm{k}}) (\mathbf{G}_C \bar{\mathbf{H}}^{\infty\dagger}_{L,C} \mathbf{c}_{L,\mathrm{k}})^\dagger \\
&= \mathbf{G}_C \bar{\mathbf{H}}^{\infty\dagger}_{L,C} \Big(2\pi \sum_{\mathrm{k}} \mathbf{c}_{L,\mathrm{k}} \delta(E - E_{\mathrm{k}}) \mathbf{c}^\dagger_{L,\mathrm{k}} \Big) \bar{\mathbf{H}}^{\infty}_{L,C} \mathbf{G}^\dagger_C \\
&= \mathbf{G}_C \big(\bar{\mathbf{H}}^{\infty\dagger}_{L,C} \mathbf{G}^r_L \bar{\mathbf{H}}^{\infty}_{L,C} - \bar{\mathbf{H}}^{\infty\dagger}_{L,C} \mathbf{G}^a_L \bar{\mathbf{H}}^{\infty}_{L,C} \big) \mathbf{G}^\dagger_C \\
&= \mathbf{G}_C \mathbf{\Gamma}_L \mathbf{G}^\dagger_C
\end{aligned} \tag{B.28}$$

for the + states coming from the left electrode, where, in the second step, we defined $\mathbf{A}_L = 2\pi \sum_{\mathrm{k}} \mathbf{c}_{L,\mathrm{k}} \delta(E - E_{\mathrm{k}}) \mathbf{c}^\dagger_{L,\mathrm{k}}$ and applied Eq. (B.22), and to obtain the final expression, we used Eqs. (B.13) and (B.11) and that $\mathbf{G}^a = (\mathbf{G}^r)^\dagger$ for real E. A similar formula is obtained for the − states coming from the right electrode, with the replacements $\bar{\mathbf{H}}^{\infty}_{L,C} \to \bar{\mathbf{H}}^{\infty\dagger}_{L,C}$ and $L \to R$ in Eq. (B.28).

An important special case of Eq. (B.27) is if $\mu_L = \mu_R \equiv \mu$, that is, when the entire system is in equilibrium. It can be shown [4, 39] that $\mathbf{A}^+ + \mathbf{A}^- = \mathbf{G}_C(\mathbf{\Gamma}_L + \mathbf{\Gamma}_R)\mathbf{G}^\dagger_C = -\mathrm{Im}\{\mathbf{G}_C\}$, which leads to the expression

$$\mathbf{n} = -\frac{2}{\pi}\int_{-\infty}^{\infty} \mathrm{Im}\{\mathbf{G}_C\} f(E-\mu)\,\mathrm{d}E \tag{B.29}$$

for the density matrix in equilibrium at the chemical potential μ.

The expressions derived above can also be written in the more convenient notation of the so-called "lesser" Green's function $\mathbf{G}^<$ [44], in which we define

$$\mathbf{G}^< = \mathbf{G}^r_C \mathbf{\Sigma}^< \mathbf{G}^a_C, \tag{B.30}$$

where (if there is no inelastic scattering)

$$\mathbf{\Sigma}^< = \hat{\imath}\Big(\mathbf{\Gamma}_L f(E-\mu_L) + \mathbf{\Gamma}_R f(E-\mu_R) \Big), \tag{B.31}$$

is the lesser self-energy matrix. Subsequently, the density matrix can be evaluated from the simple expression

$$\mathbf{n} = \frac{1}{2\pi\hat{\imath}} \int_{-\infty}^{\infty} \mathbf{G}^{<} \, \mathrm{d}E. \tag{B.32}$$

which is off course equivalent to Eq. (B.26).

B.6 Current and transmission function

In the case where the reservoirs providing the electrons to the electrodes have different chemical potentials μ_L and μ_R, a current I will flow. To obtain this current we have to take into account the time evolution of the wave function, i.e., $\hat{\imath}\hbar\frac{\partial}{\partial t}\mathbf{c}_C = \mathbf{H}\mathbf{c}_C$. It can be shown [92], by considering the probability current $\frac{\partial}{\partial t}|\mathbf{c}|^2$, which is conserved in a steady-state situation, that the flow of electrons from the left electrode into the central region, at a given energy E_k, is

$$i_{L,\mathrm{k}} = -\frac{\hat{\imath}e}{\hbar}\big(\mathbf{c}_{L,\mathrm{k}}^\dagger \mathbf{H}_{L,C}^{\infty} \mathbf{c}_{C,\mathrm{k}} - \mathbf{c}_{C,\mathrm{k}}^\dagger \mathbf{H}_{L,C}^{\infty\dagger} \mathbf{c}_{L,\mathrm{k}}\big), \tag{B.33}$$

where $i_{L,\mathrm{k}}$ is defined as positive for charge flowing into the device (we have explicitly used SI units for this formula). A similar expression is obtained for the right electrode, equal to $L \to R$ in Eq. (B.33).

Suppose that an electron is incident from the left electrode in mode $\mathbf{c}_{L,\mathrm{k}}$, which according to Eq. (B.25) gives rise to the scattering state

$$\mathbf{c}_{C,\mathrm{k}} = -\mathbf{G}_C \bar{\mathbf{H}}_{L,C}^{\infty\dagger} \mathbf{c}_{L,\mathrm{k}}, \tag{B.34}$$

in the central region and subsequently

$$\mathbf{c}_{R,\mathrm{k}} = -\mathbf{G}_R \bar{\mathbf{H}}_{R,C}^{\infty} \mathbf{c}_{C,\mathrm{k}}, \tag{B.35}$$

in the right electrode (for the latter we use arguments similar to those leading up to Eq. (B.5)). Then, the right-going (+) current arising as this electron passes all the way to the right electrode can be expressed as

$$\begin{aligned} i_\mathrm{k}^{+} &= -\frac{\hat{\imath}e}{\hbar}\big(\mathbf{c}_{R,\mathrm{k}}^\dagger \mathbf{H}_{R,C}^{\infty} \mathbf{c}_{C,\mathrm{k}} - \mathbf{c}_{C,\mathrm{k}}^\dagger \mathbf{H}_{R,C}^{\infty\dagger} \mathbf{c}_{R,\mathrm{k}}\big) \\ &= -\frac{e}{\hbar}\mathbf{c}_{C,\mathrm{k}}^\dagger \mathbf{\Gamma}_R \mathbf{c}_{C,\mathrm{k}} \\ &= -\frac{e}{\hbar}\mathbf{c}_{L,\mathrm{k}}^\dagger \bar{\mathbf{H}}_{L,C}^{\infty} \mathbf{G}_C^\dagger \mathbf{\Gamma}_R \mathbf{G}_C \bar{\mathbf{H}}_{L,C}^{\infty\dagger} \mathbf{c}_{L,\mathrm{k}}, \end{aligned} \tag{B.36}$$

by combining Eqs. (B.33)–(B.35). In the same manner, we can write the left-going (−) current of electrons originating from the right electrode, as

$$i_\mathrm{k}^{-} = -\frac{e}{\hbar}\mathbf{c}_{R,\mathrm{k}}^\dagger \bar{\mathbf{H}}_{R,C}^{\infty\dagger} \mathbf{G}_C^\dagger \mathbf{\Gamma}_L \mathbf{G}_C \bar{\mathbf{H}}_{R,C}^{\infty} \mathbf{c}_{R,\mathrm{k}}. \tag{B.37}$$

The total current I through the two-probe system is subsequently obtained by subtracting the right-going and left-going currents for all the modes, taking into account that the electrons are provided by reservoirs at chemical potentials μ_L and μ_R, i.e.,

$$I = 2\int_{-\infty}^{\infty} \left(I^{+} f(E-\mu_L) - I^{-} f(E-\mu_R)\right)\, \mathrm{d}E \tag{B.38}$$

where the factor 2 is from spin degeneracy, f is the Fermi function, and

$$\begin{aligned} I^{+} &= \frac{e}{\hbar}\sum_{\mathrm{k}} \delta(E-E_{\mathrm{k}}^{\pm}) i_{\mathrm{k}}^{+} \\ &= \frac{e}{2\pi\hbar}\mathrm{Tr}\Big\{\bar{\mathbf{H}}_{L,C}^{\infty\dagger}\Big(2\pi\sum_{\mathrm{k}} \mathbf{c}_{L,\mathrm{k}}\delta(E-E_{\mathrm{k}})\mathbf{c}_{L,\mathrm{k}}^{\dagger}\Big)\bar{\mathbf{H}}_{L,C}^{\infty}\mathbf{G}_C^{\dagger}\boldsymbol{\Gamma}_R\mathbf{G}_C\Big\} \\ &= \frac{e}{\hbar}\mathrm{Tr}\{\boldsymbol{\Gamma}_L\mathbf{G}_C^{\dagger}\boldsymbol{\Gamma}_R\mathbf{G}_C\} \end{aligned} \tag{B.39}$$

using the same arguments as for Eq. (B.28) together with the fact that $\mathrm{Tr}\{\mathbf{b}\mathbf{a}^{\dagger}\} = \mathrm{Tr}\{\mathbf{a}^{\dagger}\mathbf{b}\} = \mathbf{a}^{\dagger}\mathbf{b}$ for any vectors $\mathbf{a}$ and $\mathbf{b}$ of equal size. In a completely similar derivation, we can obtain

$$I^{-} = -\frac{e}{\hbar}\mathrm{Tr}\{\boldsymbol{\Gamma}_R\mathbf{G}_C^{\dagger}\boldsymbol{\Gamma}_L\mathbf{G}_C\} = I^{+}, \tag{B.40}$$

for the left-going current, again using the properties of the trace operator. Finally, we can rewrite the total the current formula in Eq. (B.36) in the familiar form

$$I = \frac{2e}{h}\int_{-\infty}^{\infty} T(E)\left(f(E-\mu_L) - f(E-\mu_R)\right)\, \mathrm{d}E, \tag{B.41}$$

where

$$T(E) = \mathrm{Tr}\{\boldsymbol{\Gamma}_L\mathbf{G}_C^{\dagger}\boldsymbol{\Gamma}_R\mathbf{G}_C\}, \tag{B.42}$$

is called the transmission function. This expression, which is credited to Caroli [57], can also be derived from the Landauer-Büttiker formalism [4, 55] and is widely used in methods for quantum transport.

Appendix C

Solution of linearized QEPs

Several of the methods for calculating the transmission $T(E)$ described in this thesis rely heavily on the efficient and numerically stable solution of quadratic eigenvalue problems. In this brief appendix, we show how to do this most efficiently via standard LAPACK calls. A simple trick is described, that can be characterized as a code optimization rather than a change to the theoretical model itself. It is generally applicable to all QEPs.

C.1 Shift-and-invert QEP

The simplest and currently state-of-the-art way to solve a dense QEP of size N,

$$\big(\lambda^2 \boldsymbol{M} + \lambda \boldsymbol{C} + \boldsymbol{K}\big)\boldsymbol{x} = 0, \tag{C.1}$$

is by linearizing to a generalized eigenvalue problem of size $2N$, given by [63]

$$\boldsymbol{A}\boldsymbol{z} = \lambda \boldsymbol{B}\boldsymbol{z}, \quad \boldsymbol{A} = \begin{pmatrix} 0 & \boldsymbol{I} \\ -\boldsymbol{K} & -\boldsymbol{C} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{M} \end{pmatrix}, \tag{C.2}$$

where the $2N$ eigenvalues λ are unchanged and $\boldsymbol{z}^T = (\boldsymbol{x}^T, \lambda \boldsymbol{x}^T)$ so that the eigenvectors $\boldsymbol{x}$ can be selected as the first N elements of $\boldsymbol{z}$.

Implementing this solution in practice using LAPACK routines [1] we found that it was favorable to take advantage of the known block structure of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ by performing a shift-and-invert transformation and solve

$$\tilde{\boldsymbol{A}}\boldsymbol{z} = \tilde{\lambda}\boldsymbol{z}, \quad \tilde{\boldsymbol{A}} = (\boldsymbol{A} - \sigma \boldsymbol{B})^{-1}\boldsymbol{B} \tag{C.3}$$

where an eigenpair $(\tilde{\lambda}, \boldsymbol{z})$ corresponds to an eigenpair $(\lambda = \tilde{\lambda}^{-1} + \sigma, \boldsymbol{z})$ of Eq. (C.2). The 2×2 block form of $\tilde{\boldsymbol{A}}$ can be determined using the identity

[1] We call ZGGEV (based on QZ) and ZGEEV (based on QR) to solve Eq. (C.2) and Eq. (C.3), respectively.

for the inverse of a 2×2 block matrix given in Eq. (1.2). We get

$$\begin{aligned}\tilde{\boldsymbol{A}} = (\boldsymbol{A} - \sigma\boldsymbol{B})^{-1}\boldsymbol{B} &= \begin{pmatrix} -\sigma\boldsymbol{I} & \boldsymbol{I} \\ -\boldsymbol{K} & -\boldsymbol{C} - \sigma\boldsymbol{M} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{M} \end{pmatrix} \\ &= \begin{pmatrix} -\sigma^{-1}\boldsymbol{I} - \sigma^{-2}\boldsymbol{S}^{-1}\boldsymbol{K} & \sigma^{-1}\boldsymbol{S}^{-1} \\ -\sigma^{-1}\boldsymbol{S}^{-1}\boldsymbol{K} & \boldsymbol{S}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{M} \end{pmatrix} \\ &= \begin{pmatrix} -\sigma^{-1}\boldsymbol{I} - \sigma^{-1}(\sigma\boldsymbol{S})^{-1}\boldsymbol{K} & (\sigma\boldsymbol{S})^{-1}\boldsymbol{M} \\ -(\sigma\boldsymbol{S})^{-1}\boldsymbol{K} & \sigma(\sigma\boldsymbol{S})^{-1}\boldsymbol{M} \end{pmatrix},\end{aligned} \tag{C.4}$$

where the Schur complement matrix is simply $\sigma\boldsymbol{S} = -(\sigma^2\boldsymbol{M} + \sigma\boldsymbol{C} + \boldsymbol{K})$ and assumed to be well-conditioned. Now all that is needed to setup matrix $\tilde{\boldsymbol{A}}$ instead of $\boldsymbol{A}$ and $\boldsymbol{B}$ is to solve the two equations $\boldsymbol{S}\boldsymbol{X} = \boldsymbol{K}$ and $\boldsymbol{S}\boldsymbol{X} = \boldsymbol{M}$ of size $N \times N$. The solution of Eq. (C.3) is subsequently obtained by a *standard* eigenproblem algorithm and then the eigenvalues $\tilde{\lambda}$ are back-transformed into $\lambda = \tilde{\lambda}^{-1} + \sigma$ one by one. Our profiling tests show, that in all but the smallest cases, this is faster than solving the original generalized eigenvalue problem Eq. (C.2) in spite of the additional overhead from setting up $\tilde{\boldsymbol{A}}$.

Bibliography

[1] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, October 1996.

[2] Neil W. Ashcroft and David N. Mermin. *Solid State Physics*. Brooks Cole, January 1976.

[3] Richard M. Martin. *Electronic Structure*. Cambridge University Press, Cambridge, UK, April 2004.

[4] Supriyo Datta. *Quantum Transport: Atom to Transistor*. Cambridge University Press, Cambridge, UK, 2005.

[5] Hans Henrik B. Sørensen. Correlations in many-body systems with the stochastic variational method. Master's thesis, Aarhus University, Aarhus, Denmark, 2005.

[6] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:864–871, November 1964.

[7] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, Nov 1965.

[8] W. Kohn. Nobel lecture: Electronic structure of matter-wave functions and density functionals. *Rev. Mod. Phys.*, 71(5):1253–1266, Oct 1999.

[9] W. Kohn, A. D. Becke, and R. G. Parr. Density functional theory of electronic structure. *Journal of Physical Chemistry*, 100(31):12974–12980, 1996.

[10] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45(7):566–569, Aug 1980.

[11] J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23(10):5048–5079, May 1981.

[12] D. R. Hamann, M. Schlüter, and C. Chiang. Norm-conserving pseudopotentials. *Phys. Rev. Lett.*, 43(20):1494–1497, Nov 1979.

[13] G. B. Bachelet, D. R. Hamann, and M. Schlüter. Pseudopotentials that work: From H to Pu. *Phys. Rev. B*, 26(8):4199–4228, Oct 1982.

[14] N. Troullier and José Luriaas Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43(3):1993–2006, Jan 1991.

[15] Thomas L. Beck. Real-space mesh techniques in density-functional theory. *Rev. Mod. Phys.*, 72(4):1041–1080, Oct 2000.

[16] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64(4):1045–1097, Oct 1992.

[17] José M Soler, Emilio Artacho, Julian D Gale, Alberto García, Javier Junquera, Pablo Ordejón, and Daniel Sánchez-Portal. The SIESTA method for ab initio order-N materials simulation. *Journal of Physics: Condensed Matter*, 14(11):2745–2779, 2002.

[18] W. Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.*, 76(17):3168–3171, Apr 1996.

[19] Stefan Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71(4):1085–1123, Jul 1999.

[20] Hans Henrik B. Sørensen and Dan Erik Petersen. Benchmarking the Atomistix tool kit engine (poster). Contributed to Progress in Atomic Scale Modelling of Nanotechnology Systems, University of Copenhagen, August 19, 2005, unpublished.

[21] Hendrik J. Monkhorst and James D. Pack. Special points for Brillouin-zone integrations. *Phys. Rev. B*, 13(12):5188–5192, Jun 1976.

[22] K. S. Thygesen and K. W. Jacobsen. Interference and k-point sampling in the supercell approach to phase-coherent transport. *Phys. Rev. B*, 72(3):033401, 2005.

[23] The Atomistix ToolKit 2.3 manual is available online from the company homepage: `http://www.atomistix.com/manuals/ATK_2.3/html/`.

[24] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A multigrid tutorial: Second edition.* SIAM, Philadelphia, PA, USA, 2000.

[25] E. Oran Brigham. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[26] J. Taylor. *Ab-initio Modelling of Transport in Atomic Scale Devices*. PhD thesis, McGill University, Montréal, Québec, Canada, 2000.

[27] Leonard Kleinman and D. M. Bylander. Efficacious form for model pseudopotentials. *Phys. Rev. Lett.*, 48(20):1425–1428, May 1982.

[28] `http://dirac.cnrs-orleans.fr/fsatom_wiki/PseudoPotentials`.

[29] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, Jack J. Dongarra, J. Du Croz, S. Hammarling, A. Greenbaum, A. McKenney, and D. Sorensen. *LAPACK Users' guide (third ed.)*. SIAM, Philadelphia, PA, USA, 1999.

[30] P. Pulay. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chemical Physics Letters*, 73:393–398, July 1980.

[31] R. Landauer. Spatial variation of currents and fields due to localized scatterers in metallic conduction. *IBM J. Res. Dev.*, 1:223–231, 1957.

[32] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas. Generalized many-channel conductance formula with application to small rings. *Phys. Rev. B*, 31(10):6207–6215, May 1985.

[33] K. Hirose, T. Ono, Y. Fujimoto, and S. Tsukamoto. *First-Principles Calculations in Real-Space Formalism : Electronic Configurations and Transport Properties of Nanostructures*. Imperial College Press, 2005.

[34] Yigal Meir and Ned S. Wingreen. Landauer formula for the current through an interacting electron region. *Phys. Rev. Lett.*, 68(16):2512–2515, Apr 1992.

[35] N. Agraït, A. L. Yeyati, and J. M. van Ruitenbeek. Quantum properties of atomic-sized conductors. *Phys. Rep*, 377:81–279, April 2003.

[36] S. Douglas Stone and Aaron Szafer. What is measured when you measure a resistance? The landauer formula revisited. *IBM J. Res. Dev.*, 32(3):384–413, 1988.

[37] K. Stokbro, J. Taylor, M. Brandbyge, and H. Guo. Ab-initio non-equilibrium green's function formalism for calculating electron transport in molecular devices. In G. Cuniberti, K. Richter, and G. Fagas, editors, *Lecture Notes in Physics, Berlin Springer Verlag*, volume 680 of *Lecture Notes in Physics, Berlin Springer Verlag*, pages 117–151, 2005.

[38] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov. Modeling of Nanoscale Devices. *ArXiv Condensed Matter e-prints*, October 2006.

[39] Mads Brandbyge, José-Luis Mozos, Pablo Ordejón, Jeremy Taylor, and Kurt Stokbro. Density-functional method for nonequilibrium electron transport. *Phys. Rev. B*, 65(16):165401, Mar 2002.

[40] J. Taylor, H. Guo, and J. Wang. Ab initio modeling of quantum transport properties of molecular electronic devices. *Phys. Rev. B*, 63(24):245407, Jun 2001.

[41] K. Stokbro, J.-L. Mozos, P. Ordejon, Mads Brandbyge, and J. Taylor. Theoretical study of the nonlinear conductance of di-thiol benzene coupled to au(111) surfaces via thiol and thiolate bonds. *Comp. Mat. Sci.*, 27:151, 2003.

[42] Magnus Paulsson and Mads Brandbyge. Transmission eigenchannels from nonequilibrium Green's functions. *Phys. Rev. B*, 76(11):115117, 2007.

[43] Marco Buongiorno Nardelli. Electronic transport in extended systems: Application to carbon nanotubes. *Phys. Rev. B*, 60(11):7828–7833, Sep 1999.

[44] Y. Xue. First-principles based matrix Green's function approach to molecular electronic devices: General formalism. *Chemical Physics*, 281:151–170, August 2002.

[45] M. Di Ventra, S. T. Pantelides, and N. D. Lang. First-principles calculation of transport properties of a molecular device. *Phys. Rev. Lett.*, 84(5):979–982, Jan 2000.

[46] Sergey V. Faleev, François Léonard, Derek A. Stewart, and Mark van Schilfgaarde. Ab initio tight-binding lmto method for nonequilibrium electron transport in nanosystems. *Phys. Rev. B*, 71(19):195422, 2005.

[47] Brian Larade, Jeremy Taylor, H. Mehrez, and Hong Guo. Conductance, $i - v$ curves, and negative differential resistance of carbon atomic wires. *Phys. Rev. B*, 64(7):075420, Jul 2001.

[48] Julian Velev and William Butler. On the equivalence of different techniques for evaluating the green function for a semi-infinite system using a localized basis. *J. Phys: Condens. Matter*, 16(21):R637–R657, 2004.

[49] M. P. Lopez Sancho, J. M. Lopez Sancho, J. M. L. Sancho, and J. Rubio. Highly convergent schemes for the calculation of bulk and surface Green's functions. *J. Phys. F.*, 15:851–858, April 1985.

[50] A. Umerski. Closed-form solutions to surface Green's functions. *Phys. Rev. B*, 55(8):5266–5275, Feb 1997.

[51] A. R. Rocha, V. M. García-Suárez, S. Bailey, C. Lambert, J. Ferrer, and S. Sanvito. Spin and molecular electronics in atomically generated orbital landscapes. *Phys. Rev. B*, 73(8):085414, 2006.

[52] T. Ando. Quantum point contacts in magnetic fields. *Phys. Rev. B*, 44(15):8017–8027, Oct 1991.

[53] P. S. Krstić, X.-G. Zhang, and W. H. Butler. Generalized conductance formula for the multiband tight-binding model. *Phys. Rev. B*, 66(20):205319, Nov 2002.

[54] T Shimazaki, H Maruyama, Y Asai, and K Yamashita. A theoretical study of molecular conduction. II. A Hartree-Fock approach to transmission probability. *J. Chem. Phys.*, 123:1641111, 2005.

[55] P. A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, and P. J. Kelly. Conductance calculations for quantum wires and interfaces: Mode matching and Green's functions. *Phys. Rev. B*, 72(3):035450, 2005.

[56] Ivan Rungger and Stefano Sanvito. Accurate self-energy algorithm for quasi-1d systems. *ArXiv Condensed Matter e-prints*, 2007.

[57] C. Caroli, R. Combescot, P. Nozieres, and D. Saint-James. Direct calculation of the tunneling current. *J. Phys. C: Solid St. Phys.*, 4(8):916–929, 1971.

[58] P. Schmelcher P. S. Drouvelis and P. Bastian. Parallel implementation of the recursive Green's function method. *J. Comp. Phys.*, 215:741–756, July 2006.

[59] Petr A. Khomyakov and Geert Brocks. Real-space finite-difference method for conductance calculations. *Phys. Rev. B*, 70(19):195402, 2004.

[60] Yoshitaka Fujimoto and Kikuji Hirose. First-principles treatments of electron transport properties for nanoscale junctions. *Phys. Rev. B*, 67(19):195315, May 2003.

[61] K. Xia, M. Zwierzycki, M. Talanana, P. J. Kelly, and G. E. W. Bauer. First-principles scattering matrices for spin transport. *Phys. Rev. B*, 73(6):064420, 2006.

[62] G. Brocks, V. M. Karpan, P. J. Kelly, P. A. Khomyakov, I. Marushchenko, A. Starikov, M. Talanana, I. Turek, K. Xia, P. X. Xu, M. Zwierzycki, and G. E. W. Bauer. Calculating scattering matrices by wave function matching. Ψ_k*-Newsletter*, 80:144–187, 2007.

[63] Françoise Tisseur and Karl Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43(2):235–286, 2001.

[64] Daniel S. Fisher and Patrick A. Lee. Relation between conductivity and transmission matrix. *Phys. Rev. B*, 23(12):6851–6854, Jun 1981.

[65] N. D. Lang and Ph. Avouris. Carbon-atom wires: Charge-transfer doping, voltage drop, and the effect of distortions. *Phys. Rev. Lett.*, 84(2):358–361, Jan 2000.

[66] Pawel Pomorski, Christopher Roland, and Hong Guo. Quantum transport through short semiconducting nanotubes: A complex band structure analysis. *Phys. Rev. B*, 70(11):115408, 2004.

[67] Morten Stilling, Kurt Stokbro, and Karsten Flensberg. Crystalline magnetotunnel junctions: Fe-MgO-Fe, Fe-FeOMgO-Fe and Fe-AuMgOAu-Fe. In *NSTI Nanotech 2006 Technical Proceedings*, volume 3, page 39, 2006.

[68] H. S. Gokturk. Electrical properties of ideal carbon nanotubes. In *Nanotechnology, 2005. 5th IEEE Conference on*, volume 2, pages 677–680, 2005.

[69] Ronald B. Morgan and Min Zeng. A harmonic restarted Arnoldi algorithm for calculating eigenvalues and determining multiplicity. *Linear Algebra Appl.*, 415:96, 2006.

[70] Karl Meerbergen and Dirk Roose. Matrix transformations for computing rightmost eigenvalues of large sparse non-symmetric eigenvalue problems. *IMA J. Numer. Anal.*, 16(3):297–346, 1996.

[71] M. N. Kooper, H. A. van der Vorst, S. Poedts, and J. P. Goedbloed. Application of the implicitly updated Arnoldi method with a complex shift-and-invert strategy in mhd. *J. Comput. Phys.*, 118(2):320–328, 1995.

[72] Narinder Nayar and James M. Ortega. Computation of selected eigenvalues of generalized eigenvalue problems. *J. Comput. Phys.*, 108(1):8–14, 1993.

[73] H. Voss. A Jacobi-Davidson method for nonlinear and nonsymmetric eigenproblems. *Comput. Struct.*, 85(17-18):1284–1292, 2007.

[74] U. B. Holz, G. H. Golub, and K. H. Law. A subspace approximation method for the quadratic eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 26(2):498–521, 2004.

[75] Qiang Ye. An iterated shift-and-invert Arnoldi algorithm for quadratic matrix eigenvalue problems. *Appl. Math. and Comp.*, 172:818, 2006.

[76] Leonard Hoffnung, Ren-Cang Li, and Qiang Ye. Krylov type subspace methods for matrix polynomials. *Linear Algebra Appl.*, 415:52, 2006.

[77] Zhaojun Bai and Yangfeng Su. SOAR: A second-order Arnoldi method for the solution of the quadratic eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 26(3):640–659, 2005.

[78] Lin-Wang Wang and Alex Zunger. Solving Schrödinger's equation around a desired energy: Application to silicon quantum dots. *Journal of Chemical Physics*, 100(3):2394–2397, 1994.

[79] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. van der Vorst. Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT Numerical Mathematics*, 36:595–633, 1996.

[80] Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst. *Templates for the solution of algebraic eigenvalue problems: A practical guide*. SIAM, Philadelphia, PA, USA, 2000.

[81] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math*, 9:17, 1951.

[82] Y. Saad. Variations on Arnoldi's method for computing eigen elements of large unsymmetric matrices. *Linear Algebra Appl.*, 34:269–295, 1980.

[83] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM Philadelphia, June 1997.

[84] Beresford N. Parlett and Youcef Saad. Complex shift and invert strategies for real matrices. *Linear Algebra Appl.*, 88-89:575–596, 1987.

[85] Michiel E. Hochstenbach and Henk A. van der Vorst. Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem. *SIAM J. Sci. Comput.*, 25(2):591–603, 2003.

[86] Z. Jia. Arnoldi type algorithms for large unsymmetric multiple eigenvalue problems. *J. Comp. Math.*, 17:257, 1999.

[87] M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour. Conductance of a molecular junction. *Science*, 278(5336):252–254, 1997.

[88] Abraham Nitzan and Mark A. Ratner. Electron Transport in molecular wire junctions. *Science*, 300(5624):1384–1389, 2003.

[89] J. Appenzeller, Y.-M. Lin, J. Knoch, and Ph. Avouris. Band-to-band tunneling in carbon nanotube field-effect transistors. *Phys. Rev. Lett.*, 93(19):196805, Nov 2004.

[90] G. W. Stewart. *Matrix Algorithms*. SIAM, Philadelphia, PA, USA, 2001.

[91] S. Datta. *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, Cambridge, UK, May 1997.

[92] Magnus Paulsson, Ferdows Zahid, and Supriyo Datta. Resistance of a molecule. *ArXiv Condensed Matter e-prints*, 2002.

PAPER I

Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 227 (2008) 3174–3190
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Block tridiagonal matrix inversion and fast transmission calculations

Dan Erik Petersen [a,*], Hans Henrik B. Sørensen [b], Per Christian Hansen [b], Stig Skelboe [a], Kurt Stokbro [c]

[a] Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark
[b] Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Bldg. 321, DK-2800 Lyngby, Denmark
[c] Nanoscience Center, University of Copenhagen, Universitetsparken 5d, DK-2100 Copenhagen, Denmark

Received 24 April 2007; received in revised form 15 November 2007; accepted 19 November 2007
Available online 8 December 2007

Abstract

A method for the inversion of block tridiagonal matrices encountered in electronic structure calculations is developed, with the goal of efficiently determining the matrices involved in the Fisher–Lee relation for the calculation of electron transmission coefficients. The new method leads to faster transmission calculations compared to traditional methods, as well as freedom in choosing alternate Green's function matrix blocks for transmission calculations. The new method also lends itself to calculation of the tridiagonal part of the Green's function matrix. The effect of inaccuracies in the electrode self-energies on the transmission coefficient is analyzed and reveals that the new algorithm is potentially more stable towards such inaccuracies.

PACS: 71.15.−m; 02.70.−c

Keywords: Matrix inversion; Electron transport; Transmission; Density functional theory

1. Introduction

Quantum transport simulations have become an important theoretical tool for investigating the electrical properties of nanoscale systems, both in the semi-empirical approach [1–4] and full ab initio approach [5–8]. The basis for the approach is the Landauer–Büttiker model of coherent transport, where the electrical properties of a nanoscale constriction is described by the transmission coefficients of a number of one-electron states propagating coherently through the constriction. The approach has been used successfully to describe the electrical properties of a wide range of nanoscale systems, including atomic wires, molecules and interfaces

* Corresponding author. Tel.: +45 35 32 14 00; fax: +45 35 32 14 01.
E-mail address: danerik@diku.dk (D.E. Petersen).

0021-9991/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2007.11.035

[9–18]. In order to apply the method to semiconductor device simulation, it is necessary to handle systems comprising millions of atoms, and this will require new, efficient algorithms for calculating the transmission coefficient.

In this paper, ideas and calculations behind an algorithm that provides an improvement over a widely popular technique employed in the calculation of transmission coefficient of so-called two-probe systems [15] is presented. A two-probe system consists of three regions: a left electrode region, a central scattering region and a right electrode region. The electrode regions are semi-infinite periodic systems, and the scattering region connects the two electrode regions. A one-electron tight-binding Hamiltonian is used to describe the electronic structure of the system. The tight-binding Hamiltonian can be obtained from a semi-empirical tight-binding description as obtained from an extended Hückel model [19] or through a first-principles approach as obtained when using a self-consistent density-functional Kohn–Sham Hamiltonian [20].

In the pursuit of determining the electronic structure of molecules, bulk crystals and two-probe systems, associated self-consistent DFT calculations, relevant Green's functions and ultimately calculation of the transmission of two-probe systems all involve the problem of matrix inversion in some form or another. This paper deals with matrices of a *block tridiagonal* form, which lie at the center of the problems to be solved. Block matrices will be denoted with uppercase bold letters, while lower case bold letters refer to sub-block matrices of their uppercase counterparts.

Throughout this paper, it is assumed that block tridiagonal matrix, $\mathbf{A}$, is dealt with and that it is to be inverted in order to obtain the Green's function matrix (or a part thereof). In the process of finding the Green's function matrix $\mathbf{G} = \mathbf{A}^{-1}$ that enters in DFT theory, the following equation sets up the problem [21]:

$$\mathbf{A} = \varepsilon\mathbf{S} - \mathbf{H} - \mathbf{\Sigma}^{\mathrm{L}} - \mathbf{\Sigma}^{\mathrm{R}}. \tag{1}$$

In the above expression $\mathbf{S}$ is an overlap matrix, $\mathbf{H}$ is the Hamiltonian of the system and $\mathbf{\Sigma}^{\mathrm{L}}$ and $\mathbf{\Sigma}^{\mathrm{R}}$ are the self-energies from the *left* and *right* semi-infinite electrodes, respectively. Furthermore, the matrix $\mathbf{G}$ depends on the variable ε that dictates the energy of an incoming one-electron coherent wave for which it is desired to investigate the transmission through the system. The methods developed in this paper are designed for a fixed value of ε.

The individual blocks of the matrix $\mathbf{A}$ are denoted $\mathbf{a}_{ij}$ and are assumed to be dense, complex matrices along the tridiagonal. The diagonal blocks are square matrices, while the off-diagonal blocks are typically rectangular. The structure of $\mathbf{A}$ for two relevant cases is shown in Figs. 1 and 2.

A method to obtain the Green's function matrix $\mathbf{G}$ is now devised, much in the same spirit as [22]. In order to do so, the matrix to be inverted, $\mathbf{A}$, is augmented with the identity matrix, $\mathbf{I}$.

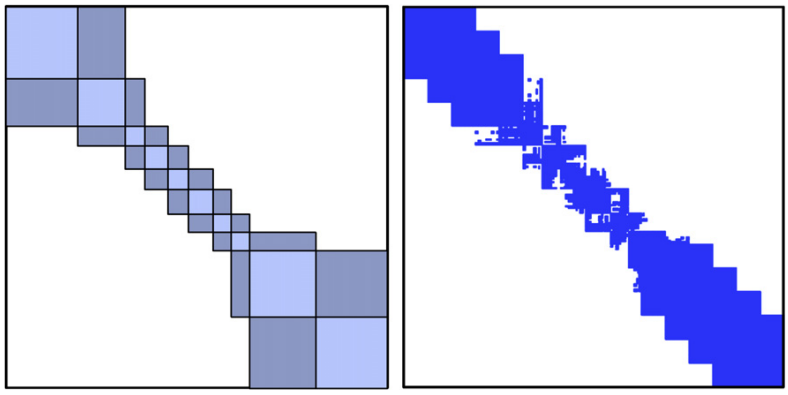


Fig. 1. The block-tridiagonal and sparsity structure for the Au111–AR example [17]. The matrix is of dimension 1295×1295, split up along the diagonal in blocks of order 243, 162, 66, 79, 69, 84, 62, 62, 225, and 243 from upper left to lower right, along with corresponding off-diagonal blocks.

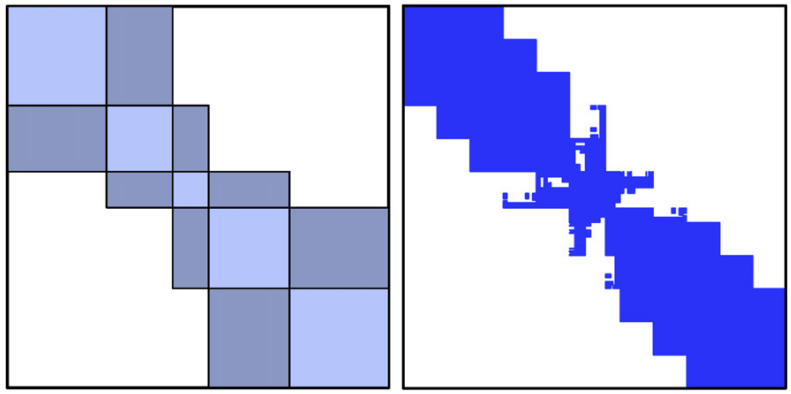


Fig. 2. The block-tridiagonal and sparsity structure for the Au111–DTB example [18]. The matrix is of dimension 943×943, split up along the diagonal in blocks of order 243, 162, 88, 198 and 243 from upper left to lower right, along with corresponding off-diagonal blocks.

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{I} \end{array}\right] = \left(\begin{array}{ccccc|ccc} \mathbf{a}_{11} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{a}_{23} & & & & \mathbf{i}_{22} & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} & & & & \mathbf{i}_{33} \\ & & \ddots & \ddots & \ddots & & & \ddots \end{array}\right) \tag{2}$$

Each diagonal block of the identity matrix, $\mathbf{i}_{ii}$ has the same square block size of the corresponding block $\mathbf{a}_{ii}$ of the matrix $\mathbf{A}$, and are themselves identity matrices.

The organization and shape of the matrix blocks in $\mathbf{A}$ are related to the topology of the two probe system. Looking at, e.g. Fig. 1, portions of the electrodes can be identified as the regions comprised of larger blocks towards the corner of the matrix, while the more sparsely populated central region of the system is identified as a series of smaller matrix blocks in the center of $\mathbf{A}$. The top left corner of $\mathbf{A}$ attaches to the left electrode, while the lower right corner attaches to the right electrode.

The expression *augmented matrix* $[\mathbf{A}|\mathbf{I}]$ is equivalent to the equation $\mathbf{AG} = \mathbf{I}$ (cf. [23]). By manipulating the augmented matrix through a series of operations such that the left side, $\mathbf{A}$, is reduced to the identity matrix $\mathbf{I}$, we will obtain the augmented matrix $[\mathbf{I}|\mathbf{G}]$ where the inverse of $\mathbf{A}$, namely $\mathbf{G} = \mathbf{A}^{-1}$, can be read on the right. This is done by illustrating the forward and backward block Gaussian elimination steps, and then combining the results.

Calculating all of $\mathbf{G}$ is ultimately not of interest. Only a block $\mathbf{g}_{ij}$ of $\mathbf{G}$ to be used in further transmission calculations will be determined. It is the particular choice of $\mathbf{g}_{ij}$ and the procedure for its calculation that separates the new transmission calculation method from previous algorithms.

This paper is organized as follows. The notation and block Gaussian elimination technique on which the methods used in this paper is based on is described in Section 2. Section 3 shows how the result of block Gaussian elimination is used to generate the Green's function matrix $\mathbf{G}$. In Section 4, the calculation of transmission values via a traditional method and a new method is explained. The new method is then benchmarked against the traditional, baseline method, via a consideration of computational complexity, as well as measured speedup times in Section 5. The effects of perturbed surface Green's function matrices on the transmission accuracy, and conclusions on which portions of $\mathbf{G}$ would lead to more accurate transmission calculations is considered in Section 6. Conclusions are finally presented in Section 7.

2. Forward and backward block Gaussian elimination

The forward procedure is characterized with the superscript L since the elimination procedure proceeds from the *left* electrode towards the right.

A block Gaussian elimination step is performed on the matrix given in Eq. (2) by multiplying the first block row by the matrix $\mathbf{c}_1^{\mathrm{L}} = -\mathbf{a}_{21}\mathbf{a}_{11}^{-1}$ and subsequently adding it to the second block row. This produces a zero block in the (2,1) position:

$$\left(\begin{array}{ccccc|ccccc} \mathbf{a}_{11} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & & & \\ \mathbf{a}_{21}+\mathbf{c}_1^{\mathrm{L}}\mathbf{a}_{11} & \mathbf{a}_{22}+\mathbf{c}_1^{\mathrm{L}}\mathbf{a}_{12} & \mathbf{a}_{23} & & & \mathbf{c}_1^{\mathrm{L}}\mathbf{i}_{11} & \mathbf{i}_{22} & & & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} & & & & \mathbf{i}_{33} & & \\ & & \ddots & \ddots & \ddots & & & & \ddots & \end{array}\right) = \left(\begin{array}{ccccc|ccccc} \mathbf{a}_{11} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & & & \\ 0 & \mathbf{a}_{22}-\mathbf{a}_{21}\mathbf{a}_{11}^{-1}\mathbf{a}_{12} & \mathbf{a}_{23} & & & \mathbf{a}_{21}\mathbf{a}_{11}^{-1}\mathbf{i}_{11} & \mathbf{i}_{22} & & & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} & & & & \mathbf{i}_{33} & & \\ & & \ddots & \ddots & \ddots & & & & \ddots & \end{array}\right). \tag{3}$$

Next, a block Gaussian elimination step is performed by multiplying the second row by the factor $\mathbf{c}_2^{\mathrm{L}} = -\mathbf{a}_{32}(\mathbf{a}_{22}-\mathbf{a}_{21}\mathbf{a}_{11}^{-1}\mathbf{a}_{12})^{-1}$ and subsequently adding it to the third row. This produces a zero block in the (3,2) position. A recursive routine that will complete a full forward block Gaussian elimination is now defined.

$$\begin{array}{ll} \mathbf{d}_{11}^{\mathrm{L}} = \mathbf{a}_{11} & \mathbf{c}_1^{\mathrm{L}} = -\mathbf{a}_{21}(\mathbf{d}_{11}^{\mathrm{L}})^{-1} \\ \mathbf{d}_{22}^{\mathrm{L}} = \mathbf{a}_{22}-\mathbf{a}_{21}(\mathbf{d}_{11}^{\mathrm{L}})^{-1}\mathbf{a}_{12} & \mathbf{c}_2^{\mathrm{L}} = -\mathbf{a}_{32}(\mathbf{d}_{22}^{\mathrm{L}})^{-1} \\ \mathbf{d}_{33}^{\mathrm{L}} = \mathbf{a}_{33}-\mathbf{a}_{32}(\mathbf{d}_{22}^{\mathrm{L}})^{-1}\mathbf{a}_{23} & \mathbf{c}_3^{\mathrm{L}} = -\mathbf{a}_{43}(\mathbf{d}_{33}^{\mathrm{L}})^{-1} \\ \quad\vdots & \quad\vdots \\ \mathbf{d}_{ii}^{\mathrm{L}} = \mathbf{a}_{ii}-\mathbf{a}_{i,i-1}(\mathbf{d}_{i-1,i-1}^{\mathrm{L}})^{-1}\mathbf{a}_{i-1,i} & \mathbf{c}_i^{\mathrm{L}} = -\mathbf{a}_{i+1,i}(\mathbf{d}_{ii}^{\mathrm{L}})^{-1} \\ \quad\vdots & \quad\vdots \\ \mathbf{d}_{nn}^{\mathrm{L}} = \mathbf{a}_{nn}-\mathbf{a}_{n,n-1}(\mathbf{d}_{n-1,n-1}^{\mathrm{L}})^{-1}\mathbf{a}_{n-1,n} & \mathbf{c}_{n-1}^{\mathrm{L}} = -\mathbf{a}_{n,n-1}(\mathbf{d}_{n-1,n-1}^{\mathrm{L}})^{-1} \end{array}$$

The matrices $\mathbf{d}_{ii}^{\mathrm{L}}$ are the diagonal blocks of the resulting matrix on the left. It can be seen that each diagonal block is calculated from the following relation:

$$\mathbf{d}_{ii}^{\mathrm{L}} = \mathbf{a}_{ii}+\mathbf{c}_{i-1}^{\mathrm{L}}\mathbf{a}_{i-1,i}, \quad \text{where } i = 2,3,\ldots,n \text{ and } \mathbf{d}_{11}^{\mathrm{L}} = \mathbf{a}_{11}, \tag{4}$$

and each row multiplication factor is:

$$\mathbf{c}_i^{\mathrm{L}} = -\mathbf{a}_{i+1,i}(\mathbf{d}_{ii}^{\mathrm{L}})^{-1}, \quad \text{where } i = 1,2,\ldots,n-1. \tag{5}$$

The similar backward procedure is characterized with the superscript R since the elimination procedure moves from the *right* electrode towards the left. The derivation of the backwards recursive expressions follows that of the forward elimination. Each diagonal block can be calculated from the following relation:

$$\mathbf{d}_{ii}^{\mathrm{R}} = \mathbf{a}_{ii}+\mathbf{c}_{i+1}^{\mathrm{R}}\mathbf{a}_{i+1,i}, \quad \text{where } i = n-1,\ldots,2,1 \text{ and } \mathbf{d}_{nn}^{\mathrm{R}} = \mathbf{a}_{nn}, \tag{6}$$

and each row multiplication factor is:

$$\mathbf{c}_i^{\mathrm{R}} = -\mathbf{a}_{i-1,i}(\mathbf{d}_{ii}^{\mathrm{R}})^{-1}, \quad \text{where } i = n, \ldots, 3, 2. \tag{7}$$

3. Combining the two procedures

After a complete forward and backward block Gaussian elimination sweep, the augmented matrices, named $[\mathbf{D}^{\mathrm{L}}|\mathbf{J}^{\mathrm{L}}]$ and $[\mathbf{D}^{\mathrm{R}}|\mathbf{J}^{\mathrm{R}}]$, respectively, will look as follows where the matrices $\mathbf{J}^{\mathrm{L}}$ and $\mathbf{J}^{\mathrm{R}}$ are lower and upper block triangular, respectively:

$$[\mathbf{D}^{\mathrm{L}} \mid \mathbf{J}^{\mathrm{L}}] = \left(\begin{array}{ccccc|ccccc} \mathbf{d}_{11}^{\mathrm{L}} & \mathbf{a}_{12} & & & & \mathbf{i}_{11} & & & & \\ \mathbf{0} & \mathbf{d}_{22}^{\mathrm{L}} & \mathbf{a}_{23} & & & \mathbf{c}_1^{\mathrm{L}}\mathbf{i}_{11} & \mathbf{i}_{22} & & & \\ & \mathbf{0} & \mathbf{d}_{33}^{\mathrm{L}} & \mathbf{a}_{34} & & \mathbf{c}_{2,1}^{\mathrm{L}}\mathbf{i}_{11} & \mathbf{c}_2^{\mathrm{L}}\mathbf{i}_{22} & \mathbf{i}_{33} & & \\ & & \ddots & \ddots & \ddots & \vdots & \vdots & & \ddots & \end{array}\right), \tag{8}$$

$$[\mathbf{D}^{\mathrm{R}} \mid \mathbf{J}^{\mathrm{R}}] = \left(\begin{array}{ccccc|ccccc} \mathbf{d}_{11}^{\mathrm{R}} & \mathbf{0} & & & & \mathbf{i}_{11} & \mathbf{c}_2^{\mathrm{R}}\mathbf{i}_{22} & \mathbf{c}_{2,3}^{\mathrm{R}}\mathbf{i}_{33} & \cdots & \\ \mathbf{a}_{21} & \mathbf{d}_{22}^{\mathrm{R}} & \mathbf{0} & & & & \mathbf{i}_{22} & \mathbf{c}_3^{\mathrm{R}}\mathbf{i}_{33} & \cdots & \\ & \mathbf{a}_{32} & \mathbf{d}_{33}^{\mathrm{R}} & \mathbf{0} & & & & \mathbf{i}_{33} & \cdots & \\ & & \ddots & \ddots & \ddots & & & & \ddots & \end{array}\right). \tag{9}$$

Here, the following notation was introduced:

$$\left.\begin{array}{l} \mathbf{c}_1^{\mathrm{R}}\mathbf{c}_2^{\mathrm{R}}\cdots\mathbf{c}_i^{\mathrm{R}} = \mathbf{c}_{1,2,\ldots,i}^{\mathrm{R}} \\ \mathbf{c}_i^{\mathrm{L}}\mathbf{c}_{i-1}^{\mathrm{L}}\cdots\mathbf{c}_1^{\mathrm{L}} = \mathbf{c}_{i,i-1,\ldots,1}^{\mathrm{L}} \end{array}\right\} \quad \text{where } i = 1, 2, \ldots, n. \tag{10}$$

Combining the results obtained from Eqs. (2), (8), and (9) by employing the fact that

$$\mathbf{AG} = \mathbf{I}, \quad \mathbf{D}^{\mathrm{L}}\mathbf{G} = \mathbf{J}^{\mathrm{L}}, \quad \mathbf{D}^{\mathrm{R}}\mathbf{G} = \mathbf{J}^{\mathrm{R}}, \tag{11}$$

the expression

$$(\mathbf{A} - \mathbf{D}^{\mathrm{L}} - \mathbf{D}^{\mathrm{R}})\mathbf{G} = \mathbf{I} - \mathbf{J}^{\mathrm{L}} - \mathbf{J}^{\mathrm{R}} \tag{12}$$

is examined, which can be viewed as the following augmented matrix expression:

$$\left[\begin{array}{c|c} \mathbf{B} & \mathbf{F} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A} & \mathbf{I} \end{array}\right] - \left[\begin{array}{c|c} \mathbf{D}^{\mathrm{L}} & \mathbf{J}^{\mathrm{L}} \end{array}\right] - \left[\begin{array}{c|c} \mathbf{D}^{\mathrm{R}} & \mathbf{J}^{\mathrm{R}} \end{array}\right], \tag{13}$$

where

$$\mathbf{B} = \begin{pmatrix} \mathbf{a}_{11} - \mathbf{d}_{11}^{\mathrm{L}} - \mathbf{d}_{11}^{\mathrm{R}} & & & \\ & \mathbf{a}_{22} - \mathbf{d}_{22}^{\mathrm{L}} - \mathbf{d}_{22}^{\mathrm{R}} & & \\ & & \mathbf{a}_{33} - \mathbf{d}_{33}^{\mathrm{L}} - \mathbf{d}_{33}^{\mathrm{R}} & \\ & & & \ddots \end{pmatrix} \tag{14}$$

and

$$\mathbf{F} = \begin{pmatrix} -\mathbf{i}_{11} & -\mathbf{c}_{2}^{\mathrm{R}} & -\mathbf{c}_{2,3}^{\mathrm{R}} & -\mathbf{c}_{2,3,4}^{\mathrm{R}} & \cdots \\ -\mathbf{c}_{1}^{\mathrm{L}} & -\mathbf{i}_{22} & -\mathbf{c}_{3}^{\mathrm{R}} & -\mathbf{c}_{3,4}^{\mathrm{R}} & \cdots \\ -\mathbf{c}_{2,1}^{\mathrm{L}} & -\mathbf{c}_{2}^{\mathrm{L}} & -\mathbf{i}_{33} & -\mathbf{c}_{4}^{\mathrm{R}} & \cdots \\ -\mathbf{c}_{3,2,1}^{\mathrm{L}} & -\mathbf{c}_{3,2}^{\mathrm{L}} & -\mathbf{c}_{3}^{\mathrm{L}} & -\mathbf{i}_{44} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{15}$$

When $\mathbf{B}$ is subsequently reduced to the identity matrix $\mathbf{I}$, $\mathbf{F}$ will simultaneously be transformed into the Green's function matrix $\mathbf{G}$. In other words, the Green's function matrix sought for can be expressed as $\mathbf{G} = \mathbf{B}^{-1}\mathbf{F}$. The Green's function matrix is:

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_{11} & \mathbf{g}_{11}\mathbf{c}_{2}^{\mathrm{R}} & \mathbf{g}_{11}\mathbf{c}_{2,3}^{\mathrm{R}} & \cdots & \mathbf{g}_{11}\mathbf{c}_{2,\ldots,n}^{\mathrm{R}} \\ \mathbf{g}_{22}\mathbf{c}_{1}^{\mathrm{L}} & \mathbf{g}_{22} & \mathbf{g}_{22}\mathbf{c}_{3}^{\mathrm{R}} & \cdots & \mathbf{g}_{22}\mathbf{c}_{3,\ldots,n}^{\mathrm{R}} \\ \mathbf{g}_{33}\mathbf{c}_{2,1}^{\mathrm{L}} & \mathbf{g}_{33}\mathbf{c}_{2}^{\mathrm{L}} & \mathbf{g}_{33} & \cdots & \mathbf{g}_{33}\mathbf{c}_{4,\ldots,n}^{\mathrm{R}} \\ \mathbf{g}_{44}\mathbf{c}_{3,2,1}^{\mathrm{L}} & \mathbf{g}_{44}\mathbf{c}_{3,2}^{\mathrm{L}} & \mathbf{g}_{44}\mathbf{c}_{3}^{\mathrm{L}} & \cdots & \mathbf{g}_{44}\mathbf{c}_{5,\ldots,n}^{\mathrm{R}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{nn}\mathbf{c}_{n-1,\ldots,1}^{\mathrm{L}} & \mathbf{g}_{nn}\mathbf{c}_{n-1,\ldots,2}^{\mathrm{L}} & \mathbf{g}_{nn}\mathbf{c}_{n-1,\ldots,3}^{\mathrm{L}} & \cdots & \mathbf{g}_{nn} \end{pmatrix}, \tag{16}$$

where the following expression for the diagonal blocks of the Green's function matrix is introduced:

$$\mathbf{g}_{ii} = -\mathbf{b}_{ii}^{-1} = \left(-\mathbf{a}_{ii} + \mathbf{d}_{ii}^{\mathrm{L}} + \mathbf{d}_{ii}^{\mathrm{R}}\right)^{-1} \quad \text{where } i = 1, 2, \ldots, n. \tag{17}$$

Off-diagonal entries are then calculated via appropriate multiplications with calculated diagonal block matrices and factors obtained during block Gaussian elimination as follows using the notation given in Eq. (10):

$$\mathbf{g}_{ij} = \mathbf{g}_{ii}\mathbf{c}_{i+1,i+2,\ldots,j-1,j}^{\mathrm{R}} \quad \text{for } i < j \tag{18}$$

$$\mathbf{g}_{ij} = \mathbf{g}_{ii}\mathbf{c}_{i-1,i-2,\ldots,j+1,j}^{\mathrm{L}} \quad \text{for } i > j. \tag{19}$$

4. Computation of transmission

The calculation of transmission t, given by the following Fisher–Lee [24] relation obtained in non–equilibrium Green's function theory, can be expressed as (cf. [21,25]):

$$t = \mathrm{Tr}\{\mathbf{G}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{G}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\}. \tag{20}$$

Here 'Tr' denotes a matrix *trace* operation, and the dagger denotes Hermitian conjugation. Regarding $\mathbf{\Gamma}^{\mathrm{L}}$ and $\mathbf{\Gamma}^{\mathrm{R}}$, the superscripts indicate *left* and *right* electrode contact leads. These matrices are defined from the electrode self-energy [21]:

$$\mathbf{\Gamma}^{\mathrm{L}} = \hat{\imath}\left(\mathbf{\Sigma}^{\mathrm{L}} - \left(\mathbf{\Sigma}^{\mathrm{L}}\right)^{\dagger}\right), \quad \mathbf{\Gamma}^{\mathrm{R}} = \hat{\imath}\left(\mathbf{\Sigma}^{\mathrm{R}} - \left(\mathbf{\Sigma}^{\mathrm{R}}\right)^{\dagger}\right), \tag{21}$$

where $\hat{\imath}$ is the imaginary unit. These matrices are only non–zero in the $(1,1)$ block for $\mathbf{\Sigma}^{\mathrm{L}}$ and $\mathbf{\Gamma}^{\mathrm{L}}$, and in the (n,n) block for the case $\mathbf{\Sigma}^{\mathrm{R}}$ and $\mathbf{\Gamma}^{\mathrm{R}}$ (cf. [26–28]).

Two methods are now presented that can be used to calculate the transmission t given in Eq. (20).

4.1. Coupling method

The coupling method is by far the popular method of choice in the literature when transmission is to be calculated via the Green's function formalism (see [26–28]). The method is introduced here, and regarded as the *baseline* method to compare the new transmission calculation method to later in the paper.

In this method, the coupling between the left and right leads is calculated, and the transmission computed accordingly. This coupling is denoted as $\mathbf{g}_{n1}$, and it resides as the lowest left corner of the Green's function matrix $\mathbf{G}$. The calculation of transmission for a particular energy ε then becomes (cf. [26]):

$$t = \mathrm{Tr}\{\mathbf{g}_{n1}\boldsymbol{\gamma}^{\mathrm{L}}_{11}\mathbf{g}^{\dagger}_{n1}\boldsymbol{\gamma}^{\mathrm{R}}_{nn}\}, \tag{22}$$

where $\boldsymbol{\gamma}^{\mathrm{L}}_{11} = [\boldsymbol{\Gamma}^{\mathrm{L}}]_{11}$ and $\boldsymbol{\gamma}^{\mathrm{R}}_{nn} = [\boldsymbol{\Gamma}^{\mathrm{R}}]_{nn}$. Thus we introduce the notation $[\cdot]_{ij}$ which delivers the (i,j)-block, with respect to $\mathbf{A}$'s block structure, of the bracketed expression. The main task is to find $\mathbf{g}_{n1}$. From Eq. (16) it is seen that the expression for this matrix is:

$$\mathbf{g}_{n1} = \mathbf{g}_{nn}\mathbf{c}^{\mathrm{L}}_{n-1}\mathbf{c}^{\mathrm{L}}_{n-2}\cdots\mathbf{c}^{\mathrm{L}}_{2}\mathbf{c}^{\mathrm{L}}_{1}, \tag{23}$$

and we see that the only factors $\mathbf{c}^{\mathrm{L}}_{i}$ involved are all those computed in a downwards block Gaussian elimination sweep. The matrix $\mathbf{g}_{nn}$ in Eq. (23) can be obtained by considering the nth block from Eq. (17):

$$\mathbf{g}_{nn} = \left(-\mathbf{a}_{nn} + \mathbf{d}^{\mathrm{L}}_{nn} + \mathbf{d}^{\mathrm{R}}_{nn}\right)^{-1} = \left(\mathbf{d}^{\mathrm{L}}_{nn}\right)^{-1}, \tag{24}$$

since $\mathbf{d}^{\mathrm{R}}_{nn} = \mathbf{a}_{nn}$. This holds similarly for the first row of the Green's function matrix. From this, it is seen that the first and last diagonal blocks of the Green's function matrix correspond to the final blocks of upwards and downwards sweeps of block Gaussian elimination, respectively, in the following manner:

$$\mathbf{g}_{11} = \left(\mathbf{d}^{\mathrm{R}}_{11}\right)^{-1} \quad \text{and} \quad \mathbf{g}_{nn} = \left(\mathbf{d}^{\mathrm{L}}_{nn}\right)^{-1}. \tag{25}$$

4.2. Overlap method

A new method that seeks to compute the transmission much like the baseline coupling method, however via a different part of the Green's function matrix, is now introduced.

Here, the idea is again based on the transmission formula Eq. (20), however the matrices dealt with change from being a coupling between the leads to that of a coupling between two adjacent blocks somewhere in the center of the system. This corresponds to centering calculations around a diagonal block of $\mathbf{A}$. This will require us to calculate the Green's function for the kth block of interest, $\mathbf{g}_{kk}$.

The name of the method arises from the fact that calculation of a diagonal block involves a *sweep* of block Gauss elimination from both the upper left and lower right of $\mathbf{A}$ which will *overlap* on the block of interest.

The motivation behind this approach is to avoid the work in having to calculate an off-diagonal block of the Green's function matrix after a series of block Gaussian elimination sweeps. This amounts to $n-1$ matrix multiplications. In the new method, overhead will arise due to calculations involving the self-energies $\boldsymbol{\Sigma}^{\mathrm{L}}$ and $\boldsymbol{\Sigma}^{\mathrm{R}}$, and the corresponding matrices $\boldsymbol{\Gamma}^{\mathrm{L}}$ and $\boldsymbol{\Gamma}^{\mathrm{R}}$. However, these computations are less expensive matrix addition operations, and they are negligible with increasing number of matrix blocks and block sizes.

As we shall demonstrate below, it is advantageous to choose k corresponding to the smallest diagonal block inside the block tridiagonal matrix $\mathbf{A}$. Although this approach involves some additional computations with the self-energy matrices and their corresponding coupling matrices, this overhead is acceptable due to the savings involved in the cheaper matrix computations for the overlap method.

Choosing an arbitrary kth diagonal block, the transmission is given in the following expression, derived in the appendix:

$$t = \mathrm{Tr}\{\mathbf{g}_{kk}[\boldsymbol{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}\mathbf{g}^{\dagger}_{kk}[\boldsymbol{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}\}, \tag{26}$$

where the new self-energy related terms are given by Eqs. (48) and (49) in the Appendix. Using the nonzero structure of the respective self-energies $\sum^{\mathrm{L}}$ and $\sum^{\mathrm{R}}$ we obtain the simpler relations

$$\left[\boldsymbol{\Gamma}^{\mathrm{L}}_{\downarrow 1}\right]_{11} = \boldsymbol{\gamma}^{\mathrm{L}}_{11}, \qquad \left[\boldsymbol{\Gamma}^{\mathrm{R}}_{\uparrow 1}\right]_{11} = \hat{\imath}\left(\left(\mathbf{d}^{\mathrm{R}}_{11}\right)^{\dagger} - \mathbf{d}^{\mathrm{R}}_{11}\right) \tag{27}$$

$$\left[\boldsymbol{\Gamma}^{\mathrm{R}}_{\uparrow n}\right]_{nn} = \boldsymbol{\gamma}^{\mathrm{R}}_{nn}, \qquad \left[\boldsymbol{\Gamma}^{\mathrm{L}}_{\downarrow n}\right]_{nn} = \hat{\imath}\left(\left(\mathbf{d}^{\mathrm{L}}_{nn}\right)^{\dagger} - \mathbf{d}^{\mathrm{L}}_{nn}\right) \tag{28}$$

and for $k = 2,\ldots,n-1$

$$[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{kk})^{\dagger} - \mathbf{d}^{\mathrm{L}}_{kk}) - \boldsymbol{\gamma}^{\mathrm{R}}_{kk}, \quad [\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{R}}_{kk})^{\dagger} - \mathbf{d}^{\mathrm{R}}_{kk}) - \boldsymbol{\gamma}^{\mathrm{L}}_{kk}. \tag{29}$$

5. Benchmark results

The methods introduced here were implemented in C++ within Atomistix's Atomistix ToolKit, and computing times were obtained for calculating the transmission for 10 different energies ε for several different systems. These systems have been taken from the literature, and an overview of selected examples is presented in Table 1.

5.1. Operation count

In order to determine which transmission method may be algorithmically more efficient, the quantity of matrix factorizations, multiplications and additions related to the three different methods available is recorded in Table 2.

In Table 3 operation counts for the calculations of the full inverse of **A** as well as calculation of only the block tridiagonal part of the inverse is included. This is done for both a Gauss elimination (GE) algorithm, as well as the new method presented in this paper.

The block tridiagonal part of the inverse is of interest for further calculations carried out in Density Functional Theory (DFT) via the Green's function formalism, and results for the full inverse are included in order to show how the new method in this paper, though suited for the block tridiagonal calculation, is ill-suited to calculate the entire inverse, compared to traditional methods.

Looking at operation counts in Table 2 on obtaining various parts of the Green's function matrix **G**, it is seen that all choices require n LU factorizations, where n is the number of diagonal blocks in **A**.

Table 1
An overview of the test examples examined in this paper

Example systems				
System	Article	Order	n	Block order
`Al100+C7`	[15]	444	5	128, 72, 16, 100, 128
`AlLead+C7`	[15]	296	5	72, 72, 20, 60, 72
`Aulll-AR`	[17]	1295	10	243, 162, 66, 79, 69, 84, 62, 62, 225, 243
`Aulll-TW`	[16]	1155	8	243, 162, 62, 70, 53, 70, 252, 243
`Aulll-DTB`	[18]	928	5	243, 162, 88, 198, 243
`Fe-MgO-Fe`	[13,14]	228	5	54, 45, 30, 45, 54
`nanotube4_4`	–	576	4	128, 128, 192, 128

For each example the original paper related to the system, the dimension of the overall matrix **A**, the number of diagonal blocks n, and finally the size of each of the diagonal blocks, from the upper left of **A** down to the lower right is listed.

Table 2
This table illustrates the amount of basic operations performed in calculating different blocks of **G** via either block Gauss elimination (GE), the coupling method or overlap method

Green's function sub-block operation count				
Block	Method	LU-factorizations	Multiplications	Additions
$\mathbf{g}_{n1}$	GE	n	$3(n-1)$	$n-1$
$\mathbf{g}_{nn}$	GE	n	$2n-1$	$n-1$
$\mathbf{g}_{kk}$	GE	n	$4n-2k-1$	$2n-k-1$
$\mathbf{g}_{n1}$	Coupling	n	$3(n-1)$	$n-1$
$\mathbf{g}_{nn}$	Overlap	n	$2n-1$	$n-1$
$\mathbf{g}_{kk}$	Overlap	n	$2n-1$	$n+1$

The third, fourth and fifth columns refer to the basic matrix operations of LU-factorization, multiplication and addition. The term n is the total amount of diagonal blocks in **A**, and k indicates which diagonal block in the Green's function matrix **G** is used for transmission calculations.

Table 3
This table illustrates the amount of basic operations performed in calculating either the full inverse **G** of **A**, or only the block tridiagonal part of it, using different methods

Green's function operation count				
Calculation	Method	LU-factorizations	Multiplications	Additions
Full inv	GE	n	$2n^2 + n - 2$	$\frac{1}{2}(n^2 + n - 2)$
Trid inv	GE	n	$\frac{1}{2}(3n^2 + 5n - 6)$	$\frac{1}{2}(n^2 + n - 2)$
Full inv	Paper	$3n - 2$	$n^2 + 4n - 4$	$4n - 6$
Trid inv	Paper	$3n - 2$	$7n - 6$	$4n - 6$

The methods employed are block Gauss Elimination (*GE*), and the new method incorporating forward and backward Gaussian elimination sweeps (*paper*), as presented in Eq. (16). The third, fourth and fifth columns refer to the basic matrix operations of LU–factorization, multiplication and addition. The term *full inv* refers to calculating the full inverse, while *trid inv* refers to obtaining only the block tridiagonal part of the inverse. The term n is the total amount of diagonal blocks in **A**.

In obtaining $\mathbf{g}_{n1}$, block Gauss elimination and the coupling method both require the same amount of operations to complete, and there is no advantage either way. Again, in obtaining the lower diagonal block $\mathbf{g}_{nn}$, both block Gauss elimination and the overlap method require the same amount of matrix–matrix calculations.

The advantage of the overlap method over block Gauss elimination occurs when a central diagonal block $\mathbf{g}_{kk}$ is required. Here, only two more matrix–matrix additions over the overlap method for $\mathbf{g}_{nn}$ is needed, while for block Gauss elimination, a series of matrix–matrix multiplies and additions add up in order to back–solve up towards the desired diagonal block. Thus the overlap method is better suited for determining diagonal blocks than block Gauss elimination.

Looking at which block of the matrix **G** is cheapest to compute on the basis of Table 2, one would apparently choose $\mathbf{g}_{nn}$. This, however, may not be the case since the table does not take into account differing block sizes among the different sub-blocks in **A** and **G**. These differing sizes can lead to substantial changes in costs regarding the basic operations of LU-factorization, matrix multiplication and matrix addition in the table. The speedup results presented later in Section 5.2 and Table 4 will verify this.

With regard to the cost of the basic operations on a matrix block of order n_i, then the amount of work for each LU-factorization, multiplication and addition is on the order of $2/3n_i^3$, $2n_i^3$ and $2n_i^2$, respectively.

5.1.1. Transmission calculation

To finally calculate transmission after successfully obtaining a sub-block of **G**, the Fisher–Lee relation (cf. Eq. (20)) is invoked, and thus three matrix–matrix multiplications are incurred, as well as a matrix trace operation. However, the significant factor here among the different methods reviewed is that the final matrix block dimensions in the Fisher–Lee relation may be different. Typically, due to the topology of the two-probe system, the central region, and thus the kth diagonal block $\mathbf{g}_{kk}$, will be of smaller size than the corner blocks $\mathbf{g}_{nn}$ or $\mathbf{g}_{n1}$. Thus a significant prefactor cost in execution time can be saved by selecting the transmission method centered around the smallest Green's function diagonal matrix block.

Table 4
This table illustrates the speedup achieved by using the new methods centered around diagonal blocks, relative to the baseline coupling method using the off-diagonal block $\mathbf{g}_{n1}$

Speedup measurements			
System	Coupling – $\mathbf{g}_{n1}$	Overlap – $\mathbf{g}_{nn}$	Overlap – $\mathbf{g}_{kk}$
`AllOO+C7`	1.0000	1.2099	2.6557
`AlLead+C7`	1.0000	1.1916	2.0092
`Aulll–AR`	1.0000	1.4211	3.2121
`Aulll–TW`	1.0000	1.3721	2.8994
`Aulll–DTB`	1.0000	1.3675	3.2537
`Fe–MgO–Fe`	1.0000	1.3064	1.8001
`nanotube4_4`	1.0000	1.2261	1.2477

The expression $\mathbf{g}_{n1}$ refers to the coupling method, while $\mathbf{g}_{nn}$ and $\mathbf{g}_{kk}$ refer to the overlap method performed on the nth and smallest diagonal block, respectively. The overlap methods are always faster, and in particular those centred on the smallest, kth, diagonal block.

Some overhead arises in choosing a central diagonal block in the shape of recalculating new matrices $[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}$ and $[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}$ for the transmission function Eq. (26) via Eqs. (27)–(29), but as these operations are cheaper matrix–matrix addition operations on small matrices, this overhead is offset by the gains in being able to employ smaller matrices in the more expensive matrix–matrix multiplication operations in the Fisher–Lee relation in Eq. (26).

5.1.2. Full inversion

With regard to determining the full inverse **G** from **A**, it is seen in Table 3 how block Gauss elimination excels over the method in this paper in terms of costly LU factorizations. Although Gauss elimination has about twice the number of matrix multiplies than the new method, Gauss elimination is still preferable when taking into account that it only requires about a third LU factorizations compared to the new method. Thus the new method is not suited for determining the full matrix **G**.

However, when requiring only the tridiagonal part of the inverse, as is the case for some DFT applications, the new method is a better choice since it only requires on the order of n matrix–matrix multiplications, while block Gauss elimination still requires on the order of n^2 matrix–matrix multiplications.

5.2. Speedup results

For an overview of the speedup of the new methods relative to the baseline coupling method, see Table 4. Overall, speedup improves in every case when moving from the coupling method to the overlap method. This is not surprising, seeing how the main difference between these two methods, operation count–wise, is the lack of extra matrix multiplications in order to obtain an off-diagonal Green's function matrix block. Eliminating this task will always lead to a faster method.

Performing calculations using the smallest diagonal block k over the first or nth block can also yield significant improvements in execution time, depending on the topology of the two-probe system, and the subsequent block structure in **A**. The difference here is that it is no longer possible to 'recycle' one of the self-energy terms that is assumed to be available from the outset, as well as different block size between $\mathbf{g}_{nn}$ and $\mathbf{g}_{kk}$. Thus in seeking a smaller diagonal matrix block to work with, appropriate self-energy terms must be determined once again, and this leads to extra overhead.

However, it may pay off to select some central diagonal block over a corner diagonal block in order to calculate transmission. This comes in the form of being able to work with smaller matrices, and thus matrix operation costs decrease. Crucially, matrix sizes may decrease such that memory requirements for matrix operations can be fulfilled by lower level hardware caches, leading to significant speedup in execution time. This effect is visible in Table 5, where significant speedup is achieved in the matrix–matrix operations involved in the Fisher–Lee calculation.

Table 5
This table illustrates the speedup in the calculation of solely the Fisher–Lee relation (see Eq. (26)) achieved by using the new methods centered around diagonal blocks, relative to the baseline coupling method using the off-diagonal block $\mathbf{g}_{n1}$

Speedup measurements – Fisher–Lee				
System	Coupling – $\mathbf{g}_{n1}$	Overlap – $\mathbf{g}_{nn}$	Overlap – $\mathbf{g}_{kk}$	Theoretical – $\frac{n^3}{m^3}$
`AllOO+C7`	1.0000	1.0567	548.4500	512.000
`AlLead+C7`	1.0000	0.9546	47.4516	46.656
`Aulll-AR`	1.0000	1.2654	170.6912	60.207
`Aulll-TW`	1.0000	1.2788	275.6198	96.381
`Aulll-DTB`	1.0000	1.2716	59.8502	21.056
`Fe-MgO-Fe`	1.0000	1.3186	7.1354	5.832
`nanotube4_4`	1.0000	0.9940	1.0178	1.000

The expression $\mathbf{g}_{n1}$ refers to the coupling method, while $\mathbf{g}_{nn}$ and $\mathbf{g}_{kk}$ refer to the overlap method performed on the nth and smallest diagonal block, respectively. The final column indicates the theoretical speedup based on the $\mathcal{O}(n^3)$ cost of evaluating Eq. (26). The reason for better speedup over theoretical prediction is due to improved cache usage by the smaller matrices dealt with when using $\mathbf{g}_{kk}$.

Furthermore, as will be explored in Section 6 concerning transmission accuracy, depending on the system, central matrix blocks may be less prone to perturbation from inaccurately calculated electrode surface Green's function matrices. This, however, varies from system to system, as well as incoming electron wave energies ε.

6. Transmission accuracy

It has been shown that any block of the Green's function matrix can be used in order to calculate transmission and a new strategy employing diagonal blocks of $\mathbf{G}$ was developed. The question now is which part of $\mathbf{G}$ might be used in order to achieve best accuracy in determining transmission. This section suggests that an investigation of the accuracy achieved for a given block may lead to informed choices. The problem of the selection of which matrix block is best concerning accuracy comes from the fact that in practice the self-energies of the electrodes, $\boldsymbol{\sigma}_{11}^{\mathrm{L}}$ and $\boldsymbol{\sigma}_{nn}^{\mathrm{R}}$, are not computed exactly. This is because in the Green's function formalism approach, the surface Green's function matrices for the electrodes (and hence their corresponding self-energies) are typically determined through an iterative procedure [29] that only converges to the correct retarded Green's function matrix when a small positive imaginary perturbation is applied. This means that transmissions are calculated for a slightly perturbed matrix $\widetilde{\mathbf{A}}$, where the corner blocks $\mathbf{a}_{11}$ and $\mathbf{a}_{nn}$ are perturbed to some degree through the inexact self-energies.

The matrix $\mathbf{A}$ here will denote the case when no imaginary perturbation is used and this can be done by employing a different manner to converge the surface Green's function matrices, such as a wave function matching [30–32] approach. To investigate how this imaginary perturbation ultimately affects the Green's function matrix that transmissions are calculated with, the inverses of an unperturbed case and a perturbed case are compared. The perturbation on $\mathbf{A}$ is described as the added matrix $\mathbf{P}$, defined as zero everywhere, except the corner blocks $\mathbf{p}_{11}$ and $\mathbf{p}_{nn}$, that correspond to the corner blocks $\mathbf{a}_{11}$ and $\mathbf{a}_{nn}$, both in size and location.

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{P}, \quad \text{where} \quad \mathbf{P} = \begin{pmatrix} \mathbf{p}_{11} & & \\ & & \\ & & \mathbf{p}_{nn} \end{pmatrix} \tag{30}$$

The perturbation matrix, as seen in Eq. (30), is divided into 9 blocks, where the empty space denotes areas with elements equal to zero. In a similar manner, the inverse $\mathbf{G} = \mathbf{A}^{-1}$ is subdivided into the same block sizes.

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & & \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \mathbf{a}_{23} & \\ & \mathbf{a}_{32} & \mathbf{a}_{33} & \mathbf{a}_{34} \\ & & \ddots & \ddots & \ddots & \\ & & & & \mathbf{a}_{n-1,n} \\ & & & \mathbf{a}_{n,n-1} & \mathbf{a}_{nn} \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} \mathbf{g}_{11} & \bullet & \bullet & \bullet & \mathbf{g}_{1n} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \mathbf{g}_{n1} & \bullet & \bullet & \bullet & \mathbf{g}_{nn} \end{pmatrix} \tag{31}$$

To investigate the effect of the perturbation $\mathbf{P}$ the derivation of $\widetilde{\mathbf{G}} = \widetilde{\mathbf{A}}^{-1}$ is carried out:

$$\widetilde{\mathbf{G}} = [\mathbf{A}(\mathbf{I} + \mathbf{G}\mathbf{P})]^{-1} = (\mathbf{I} + \mathbf{G}\mathbf{P})^{-1}\mathbf{G}. \tag{32}$$

If the perturbation is assumed to be small, such that the spectral radius satisfies $\rho(\mathbf{G}\mathbf{P}) < 1$, then the first inverse term can be expressed via a geometric series.

$$\widetilde{\mathbf{G}} = (\mathbf{I} + \mathbf{GP})^{-1}\mathbf{G} = \mathbf{G} - \mathbf{GPG} + \mathbf{GPGPG} - \cdots \tag{33}$$

Thus it can be seen that the difference in the perturbed and unperturbed inverses should be dominated by the term **GPG**. If **G** is subdivided into row and column blocks, as follows, it will be possible to proceed and derive a relatively compact expression for the structure of this first order correction term.

$$\mathbf{G} = \left(\begin{array}{c|c|c} \mathbf{g}_{11} & \cdots & \mathbf{g}_{1n} \\ \hline \vdots & \ddots & \vdots \\ \hline \mathbf{g}_{n1} & \cdots & \mathbf{g}_{nn} \end{array}\right) = \left(\begin{array}{c|c|c} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{array}\right) = \left(\begin{array}{c} \mathbf{c}_1 \\ \hline \vdots \\ \hline \mathbf{c}_n \end{array}\right) \tag{34}$$

such that

$$\mathbf{b}_1 = \begin{pmatrix} \mathbf{g}_{11} \\ \vdots \\ \mathbf{g}_{n1} \end{pmatrix}, \quad \mathbf{b}_n = \begin{pmatrix} \mathbf{g}_{n1} \\ \vdots \\ \mathbf{g}_{nn} \end{pmatrix} \tag{35}$$

and

$$\mathbf{c}_1 = \begin{pmatrix} \mathbf{g}_{11} & \cdots & \mathbf{g}_{1n} \end{pmatrix}, \quad \mathbf{c}_n = \begin{pmatrix} \mathbf{g}_{n1} & \cdots & \mathbf{g}_{nn} \end{pmatrix}. \tag{36}$$

With this notation, the first order perturbation term is written as follows:

$$\mathbf{GPG} = \mathbf{b}_1\mathbf{p}_{11}\mathbf{c}_1 + \mathbf{b}_n\mathbf{p}_{nn}\mathbf{c}_n. \tag{37}$$

It can be seen how the outer-product form of this expression will yield a dense matrix **GPG**, since **G** can generally be assumed to be dense. This indicates that the correction term's effect will depend directly on the full structure of **G**, and thus no prediction can be made about the effect of the perturbation on **G**, without calculating **G** itself.

We look at the first order perturbation at block (i, j):

$$[\mathbf{GPG}]_{ij} = [\mathbf{b}_1\mathbf{p}_{11}\mathbf{c}_1 + \mathbf{b}_n\mathbf{p}_{nn}\mathbf{c}_n]_{ij} = \mathbf{g}_{i1}\mathbf{p}_{11}\mathbf{g}_{1j} + \mathbf{g}_{in}\mathbf{p}_{nn}\mathbf{g}_{nj},$$

where the element $\mathbf{g}_{in}$ describes the amplitude of an electron propagating from site i to site n in the system. For most systems, this will decay as a function of the distance between orbitals at sites i and n, and thus the error should be smallest for Green's function blocks in the center of the cell, i.e., as far as possible from the electrodes. Thus we can expect choosing central blocks in **G** should lead to more accurate transmission calculations for most systems.

6.1. Numerical example with random perturbation

The effect of a perturbation of the electrode's surface Green's function matrices on the Green's function **G** itself is here illustrated by a numerical example. The Hamiltonian matrix **H** and overlap matrix **S** associated with Au111–AR is taken, and the matrix to be inverted is constructed as

$$\mathbf{A} = \mathbf{H} - \varepsilon\mathbf{S}, \quad \text{where } \varepsilon = 1.0. \tag{38}$$

The corner blocks of **A**, namely $\mathbf{a}_{11}$ and $\mathbf{a}_{nn}$, are then perturbed with matrices $\mathbf{p}_{11}$ and $\mathbf{p}_{nn}$. The elements of $\mathbf{p}_{11}$ and $\mathbf{p}_{nn}$ are computed as:

$$\mathbf{p}_{11} \leftarrow p_{ij} = \alpha_{ij}a_{ij}, \quad \text{where } a_{ij} \in \mathbf{a}_{11}, \text{ and} \tag{39}$$

$$\mathbf{p}_{nn} \leftarrow p_{kl} = \alpha_{kl}a_{kl}, \quad \text{where } a_{kl} \in \mathbf{a}_{nn}. \tag{40}$$

where the factors α_{ij} and α_{kl} are normally distributed with zero mean and standard deviation 10^{-5}.

Fig. 3 shows the results of the average difference of 100 perturbed inversions $\widetilde{\mathbf{G}}$ compared to **G**. From this figure, it is seen that for this particular choice of system (**H** and **S**) and energy (ε), the perturbation from the iterated self-energies cause the inverse to be most inaccurate at the corner diagonal blocks. Thus, choosing the

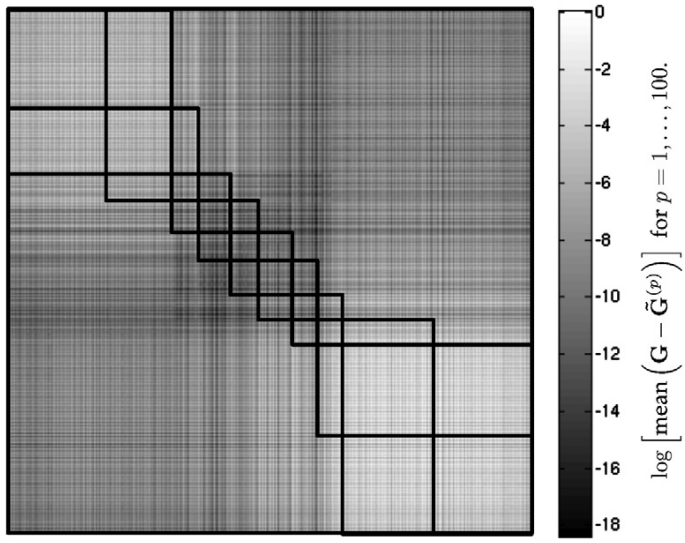


Fig. 3. The figure above illustrates the average element-wise difference expressed as $\log[\text{mean}(\mathbf{G}-\widetilde{\mathbf{G}}^{(p)})]$ for $p=1,\ldots,100$. The matrix $\mathbf{G}$ corresponds to the Au111–AR example [17]. Element-wise differences range from about the same order down to about 18 orders of magnitude smaller. The dark lines outline the original block tridiagonal structure of the original matrix $\mathbf{A}$. The logarithm employed is the base 10 logarithm. In this particular example for choice of electron energy ε and $\mathbf{A}$, the diagonal blocks in the center of the matrix suffer least in terms of accuracy.

overlap method as the transmission calculation method would be on average best served by choosing a block towards the center of the matrix, where the perturbation has the least effect. This choice is further motivated by the fact that the center blocks typically are of smaller size, and matrix operations would be faster than operations with the corner diagonal blocks.

A problem with this analogy lies in the fact that one can not predict which Green's function matrix block would provide more accurate transmission results (see Eq. (37)), without calculating the Green's function matrix in the first place. This lends prediction to be prohibitive in general, when computing transmissions. The best choice of action is then relying on the usual behavior of most two-probe systems as well as choosing the fastest calculation method, leading us to pick a diagonal block towards the center of the system, which are typically the least affected by the electrodes as well as the smallest in size.

7. Conclusion

This paper developed and introduced a new, faster method of calculating transmission for two-probe systems by using diagonal block matrices from the Green's function matrix, $\mathbf{g}_{ii}$, rather than the coupling method found extensively in the literature that uses the corner off-diagonal block $\mathbf{g}_{n1}$.

This is done by developing a method for calculating any block matrix from the Green's function matrix $\mathbf{G}$ based on a series of Gauss eliminations carried out on the original matrix $\mathbf{A}$.

To calculate transmission via a diagonal block of the Green's function matrix $\mathbf{G}$, upwards and downwards block Gaussian elimination is performed that terminates overlapping over $\mathbf{a}_{kk}$, and $\mathbf{g}_{kk}$ is calculated (cf. Eq. (17)).

Furthermore, the related coupling matrices (usually obtained via self-energy) used in the transmission formula Eq. (26) are calculated via Eqs. (27)–(29), for the new, extended electrodes. This approach dispenses with the need of a series of matrix–matrix multiplications compared to the coupling method (cf. Eq. (23)) in exchange for cheaper matrix–matrix addition operations.

Execution time measurements indicated that centering transmission calculations on the Green's function matrix's diagonal blocks was preferable, in that a series of matrix–matrix multiplications would be saved as

well as centering on smaller diagonal matrix blocks offset the cost of re-calculating self-energy matrices. Furthermore, the ability to choose smaller block matrices lends itself to the possibility of far better cache usage, and hence greater performance gains.

Perturbation analysis revealed that it is not possible to determine the effect of perturbation in the electrode self-energy matrices on the accuracy of the Green's function, without explicitly calculating the Green's function matrix. This eliminates the ability to predict which Green's function matrix block would be an ideal choice for the calculation of a two-probe system's transmission with respect to accuracy. However, due to the behavior of most two-probe systems, a central diagonal block choice is expected to yield more accurate results.

Acknowledgments

This work was supported by Grant No. 2106-04-0017, "Parallel Algorithms for Computational Nano-Science", under the NABIIT program from the Danish Council for Strategic Research.

Appendix. Derivation of Eq. (26) for the Transmission

We commence with the expression in Eq. (1). As shown in (e.g., Golub and Van Loan [33]) Section 3.2.1, we can represent a Gauss-elimination step as a matrix multiplication with a "Gauss transformation". The same is true for the block Gauss-elimination steps we use here, and thus we express a series of downwards Gauss-eliminations that terminate on row k by $\mathbf{E}_{\downarrow k}$. Similarly, a series of upwards Gauss-eliminations terminating on row k is denoted by $\mathbf{E}_{\uparrow k}$. We then write the combination of Gauss-elimination sweeps that produce a matrix $\mathbf{Z}_k$ as follows:

$$\mathbf{Z}_k = \mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}_{\uparrow k}. \tag{41}$$

Due to the structure of $\mathbf{A}$, the matrix $\mathbf{Z}_k$ is block diagonal as shown in Fig. 4. Given $\mathbf{Z}_k$, we can write the Green's function matrix as

$$\mathbf{G} = \mathbf{A}^{-1} = \mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k}. \tag{42}$$

We can then insert this expression into the Fisher–Lee relation from Eq. (20), to obtain

$$\begin{aligned} t &= \mathrm{Tr}\{(\mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k})\mathbf{\Gamma}^{\mathrm{L}}(\mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k})^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\} = \mathrm{Tr}\{\mathbf{E}_{\uparrow k}\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{E}_{\downarrow k}^{\dagger}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{E}_{\uparrow k}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\} \\ &= \mathrm{Tr}\{\mathbf{Z}_k^{-1}\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{E}_{\downarrow k}^{\dagger}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{E}_{\uparrow k}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}_{\uparrow k}\} = \mathrm{Tr}\{\mathbf{Z}_k^{-1}\mathbf{\Gamma}_{\downarrow k}^{\mathrm{L}}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{\Gamma}_{\uparrow k}^{\mathrm{R}}\} \end{aligned} \tag{43}$$

where we have introduced

$$\mathbf{\Gamma}_{\downarrow k}^{\mathrm{L}} = \mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{L}}\mathbf{E}_{\downarrow k}^{\dagger} \quad \text{and} \quad \mathbf{\Gamma}_{\uparrow k}^{\mathrm{R}} = \mathbf{E}_{\uparrow k}^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}_{\uparrow k}. \tag{44}$$

To derive Eq. (43) we used that the trace is invariant under matrix commutation [34].

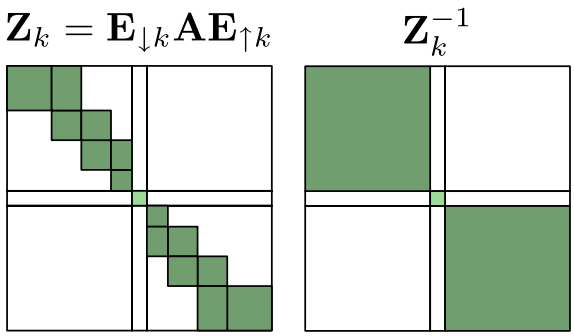


Fig. 4. The zero/nonzero structure of $\mathbf{Z}_k$ and $\mathbf{Z}_k^{-1}$.

Eq. (43) can be further simplified. First note that both $\mathbf{Z}_k$ and $\mathbf{Z}_k^{-1}$ have the special zero/nonzero structure shown in Fig. 4. Next, note that $\mathbf{\Gamma}^{\mathrm{L}}$ has nonzero elements in its (1,1)-block only, and hence the nonzeros in $\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}$ are confined to upper left blocks, as shown in Fig. 5. Similarly, the nonzeros of $\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}$ are confined to the bottom right blocks. Using the zero/nonzero structure of these matrices, it follows from the derivation illustrated in Fig. 6 that:

$$t = \mathrm{Tr}\{\mathbf{Z}_k^{-1}\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}(\mathbf{Z}_k^{-1})^{\dagger}\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}\} = \mathrm{Tr}\{[\mathbf{Z}_k^{-1}]_{kk}[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}[\mathbf{Z}_k^{-1}]^{\dagger}_{kk}[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}\}. \tag{45}$$

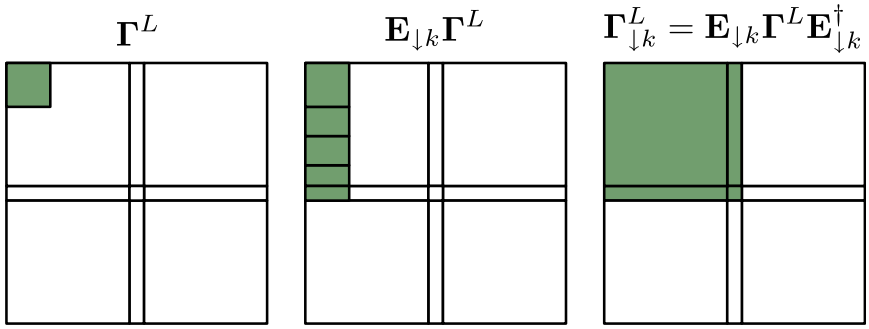


Fig. 5. The zero/nonzero structure of $\mathbf{\Gamma}^{\mathrm{L}}$ and $\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}$.

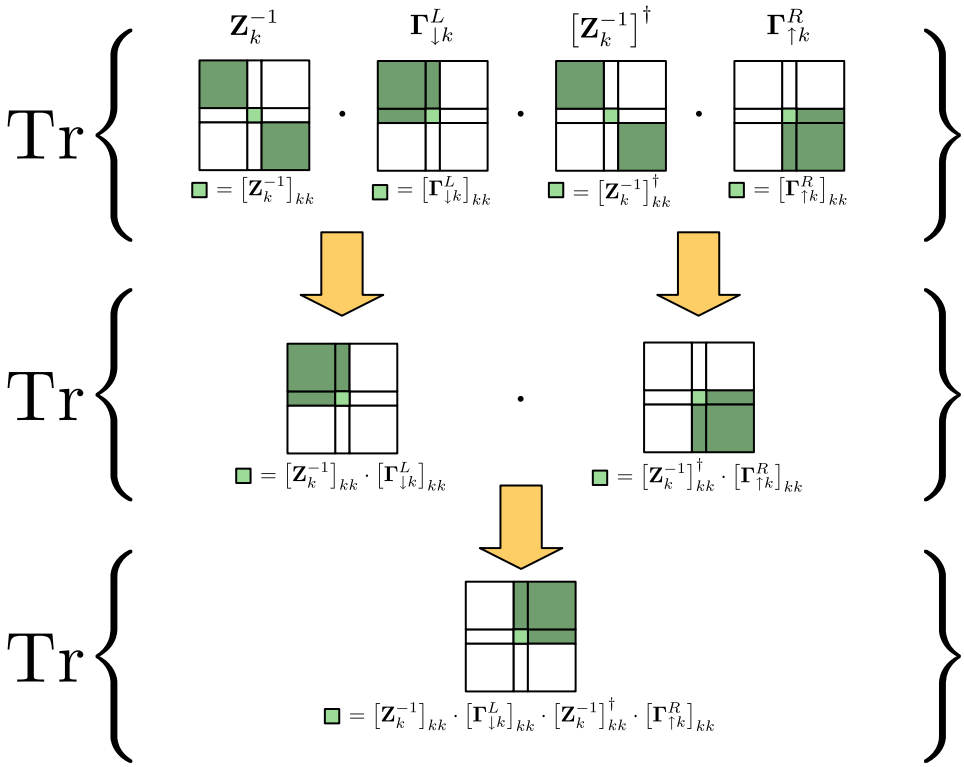


Fig. 6. Illustration of the derivation of Eq. (45) using the zero/nonzero structure of Figs. 4 and 5.

Hence, we require only the kth diagonal block of $\mathbf{Z}_k^{-1}$, and we note that this corresponds to the kth diagonal block of the Green's function matrix $\mathbf{G}$ via Eq. (17). Thus $[\mathbf{Z}_k^{-1}]_{kk} = \mathbf{g}_{kk}$, and $[\mathbf{Z}_k^{-1}]_{kk}^\dagger = \mathbf{g}_{kk}^\dagger$.

Next we consider $[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk}$ and $[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk}$. By means of Eq. (1) we can obtain the expression of a self-energy, e.g., $\mathbf{\Sigma}^{\mathrm{L}}$, and via Eq. (21) we now determine our desired matrix for the transmission calculation:

$$\begin{aligned}[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} &= [\mathbf{E}_{\downarrow k}\hat{\imath}(\mathbf{\Sigma}^{\mathrm{L}} - (\mathbf{\Sigma}^{\mathrm{L}})^\dagger)\mathbf{E}^\dagger_{\downarrow k}]_{kk} = \hat{\imath}[\mathbf{E}_{\downarrow k}((\varepsilon\mathbf{S} - \mathbf{H} - \mathbf{\Sigma}^{\mathrm{R}} - \mathbf{A}) - (\varepsilon\mathbf{S} - \mathbf{H} - \mathbf{\Sigma}^{\mathrm{R}} - \mathbf{A})^\dagger)\mathbf{E}^\dagger_{\downarrow k}]_{kk} \\ &= \hat{\imath}[\mathbf{E}_{\downarrow k}(\mathbf{A}^\dagger - \mathbf{A} - (\mathbf{\Sigma}^{\mathrm{R}} - (\mathbf{\Sigma}^{\mathrm{R}})^\dagger))\mathbf{E}^\dagger_{\downarrow k}]_{kk} \\ &= \hat{\imath}([\mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}^\dagger_{\downarrow k}]^\dagger_{kk} - [\mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}^\dagger_{\downarrow k}]_{kk}) - \hat{\imath}[\mathbf{E}_{\downarrow k}(\mathbf{\Sigma}^{\mathrm{R}} - (\mathbf{\Sigma}^{\mathrm{R}})^\dagger)\mathbf{E}^\dagger_{\downarrow k}]_{kk}. \end{aligned} \tag{46}$$

Here we used that both $\mathbf{S}$ and $\mathbf{H}$ are Hermitian and therefore vanish in the expression. The first term involving $\mathbf{A}$ is simplified via the fact that the (k,k)-subblock of the block tridiagonal $\mathbf{E}_{\downarrow k}\mathbf{A}$ remains invariant under the column operations by $\mathbf{E}^\dagger_{\downarrow k}$, and thus $[\mathbf{E}_{\downarrow k}\mathbf{A}\mathbf{E}^\dagger_{\downarrow k}]_{kk} = \mathbf{d}^{\mathrm{L}}_{kk}$. The last term, involving self-energies, is simplified via Eq. (21). We get

$$[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{kk})^\dagger - \mathbf{d}^{\mathrm{L}}_{kk}) - [\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}^\dagger_{\downarrow k}]_{kk}. \tag{47}$$

Since $\mathbf{E}_{\downarrow k}$ represents downwards elimination, the (k,k)-block in $\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}^\dagger_{\downarrow k}$ is left unaltered, i.e., $[\mathbf{E}_{\downarrow k}\mathbf{\Gamma}^{\mathrm{R}}\mathbf{E}^\dagger_{\downarrow k}]_{kk} = [\mathbf{\Gamma}^{\mathrm{R}}]_{kk} = \gamma^{\mathrm{R}}_{kk}$. Hence:

$$[\mathbf{\Gamma}^{\mathrm{L}}_{\downarrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{L}}_{kk})^\dagger - \mathbf{d}^{\mathrm{L}}_{kk}) - \gamma^{\mathrm{R}}_{kk}. \tag{48}$$

Following a similar procedure, we obtain:

$$[\mathbf{\Gamma}^{\mathrm{R}}_{\uparrow k}]_{kk} = \hat{\imath}((\mathbf{d}^{\mathrm{R}}_{kk})^\dagger - \mathbf{d}^{\mathrm{R}}_{kk}) - \gamma^{\mathrm{L}}_{kk}. \tag{49}$$

Thus we have all the terms necessary for the calculation of transmission via Eq. (43).

References

[1] P. Pernas, A. Martin-Rodero, F. Flores, Electrochemical–potential variations across a constriction, Phys. Rev. B 41 (1990) 8553–8556.
[2] W. Tian, S. Datta, Aharonov–Bohm-type effect in graphene tubules: a Landauer approach, Phys. Rev. B 49 (1994) 5097–5100.
[3] L. Chico, M. Sancho, M. Munoz, Carbon-nanotube-based quantum dot, Phys. Rev. Lett. 81 (1998) 1278–1281.
[4] A. de Parga, O.S. Hernan, R. Miranda, A.L. Yeyati, A. Martin-Rodero, F. Flores, Electron resonances in sharp tips and their role in tunneling spectroscopy, Phys. Rev. Lett. 80 (1998) 357–360.
[5] N.D. Lang, Resistance of atomic wires, Phys. Rev. B 52 (1995) 5335–5342.
[6] K. Hirose, M. Tsukada, First-principles calculation of the electronic structure for a bielectrode junction system under strong field and current, Phys. Rev. B 51 (1995) 5278–5290.
[7] M.B. Nardelli, Electronic transport in extended systems: application to carbon nanotubes, Phys. Rev. B 60 (1999) 7828–7833.
[8] M.B. Nardelli, J. Bernholc, Mechanical deformations coherent transport in carbon nanotubes, Phys. Rev. B 60 (1999) R16338–R16341.
[9] J.J. Palacios, A.J. Pérez-Jiménez, E. Louis, J.A. Vergés, Fullerene-based molecular nanobridges: a first-principles study, Phys. Rev. B 64 (2001) 115411.
[10] P.A. Derosa, J.M. Seminario, Electron transport through single molecules: scattering treatment using density functional and Green Function theories, J. Phys. Chem. B 105 (2001) 471–481.
[11] S.N. Yaliraki, A.E. Roitberg, C. Gonzalez, V. Mujica, M.A. Ratner, The injecting energy at molecule/metal interfaces: implications for conductance of molecular junctions from an ab initio molecular description, J. Chem. Phys. 111 (1999) 6997–7002.
[12] J. Taylor, H. Guo, J. Wang, Ab initio modeling of open systems: charge transfer, electron conduction, and molecular switching of a C_{60} device, Phys. Rev. B 63 (2001) 121104.
[13] M. Stilling, K. Stokbro, K. Flensberg, Electronic transport in crystalline magnetotunnel junctions: effects of structural disorder, J. Comput.-Aid. Mater. Des. 14 (2007) 141–149.
[14] M. Stilling, K. Stokbro, K. Flensberg, Crystalline magnetotunnel junctions: Fe–MgO–Fe, Fe–FeOMgO–Fe and Fe–AuMgOAu–Fe, Nanotech 3 (2006) 39–42.
[15] M. Brandbyge, J.L. Mozos, P. Ordejón, J. Taylor, K. Stokbro, Density-functional method for nonequilibrium electron transport, Phys. Rev. B 65 (2002) 165401.
[16] J. Taylor, M. Brandbyge, K. Stokbro, Theory of rectification in tour wires: the role of electrode coupling, Phys. Rev. Lett. 89 (2002) 138301.
[17] K. Stokbro, J. Taylor, M. Brandbyge, Do Aviram–Ratner diodes rectify? J. Amer. Chem. Soc. 125 (2003) 3674–3675.

[18] K. Stokbro, J.L. Mozos, P. Ordejón, M. Brandbyge, J. Taylor, Theoretical study of the nonlinear conductance of di-thiol benzene coupled to Au(1 1 1) surfaces via thiol and thiolate bonds, Comput. Mater. Sci. 27 (2003) 151–160.
[19] R. Hoffmann, An extended Hückel theory. I. Hydrocarbons, J. Chem. Phys. 39 (1963) 1397–1412.
[20] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140 (1965) A1133–A1138.
[21] S. Datta, Electronic Transport in Mesoscopic Systems, Cambridge Univ. Press, New York, 1996.
[22] E.M. Godfrin, A method to compute the inverse of an *n*-block tridiagonal quasi-Hermitian matrix, J. Phys.: Condens. Matter 3 (1991) 7843–7848.
[23] J.D. Gilbert, L. Gilbert, Linear Algebra and Matrix Theory, Academic Press Inc., 1995.
[24] D.S. Fisher, P.A. Lee, relation between conductivity and transmission matrix, Phys. Rev. B 23 (1981) 6851–6854.
[25] H. Haug, A.-P. Jauho, Quantum Kinetics in Transport and Optics of Semiconductors, Springer-Verlag, Berlin, Heidelberg, 1996.
[26] P.S. Drouvelis, P. Schmelcher, P. Bastian, Parallel implementation of the recursive Green's function method, J. Comp. Phys. 215 (2006) 741–756.
[27] S.V. Faleev, F. Léonard, D.A. Stewart, M. van Schilfgaarde, Ab initio tight-binding LMTO method for nonequilibrium electron transport in nanosystems, Phys. Rev. B 71 (2005) 195422.
[28] O. Hod, J.E. Peralta, G.E. Scuseria, First-principles electronic transport calculations in finite elongated systems: a divide and conquer approach, J. Chem. Phys. 125 (2006) 114704.
[29] M.P. López Sancho, J.M. López Sancho, J. Rubio, Highly convergent schemes for the calculation of bulk and surface Green functions, J. Phys. F: Met. Phys. 15 (1985) 851–858.
[30] H.H. Sørensen, D.E. Petersen, P.C. Hansen, S. Skelboe, K. Stokbro, Efficient wave function matching approach for quantum transport calculations, Phys. Rev. B, submitted for publication.
[31] P.A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, P.J. Kelly, Conductance calculations for quantum wires and interfaces: mode matching and Green's functions, Phys. Rev. B 72 (2005) 035450.
[32] T. Ando, Quantum point contacts in magnetic fields, Phys. Rev. B 44 (1991) 8017–8027.
[33] G.H. Golub, C.F. van Loan, Matrix Computations, Johns Hopkins University Press, London, 1996.
[34] S. Lang, Linear Algebra, Springer-Verlag, New York, 1987.

PAPER II

Efficient wave function matching approach for quantum transport calculations

Hans Henrik B. Sørensen* and Per Christian Hansen
Informatics and Mathematical Modelling, Technical University of Denmark, Bldg. 321, DK-2800 Lyngby, Denmark

Dan Erik Petersen and Stig Skelboe
Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark

Kurt Stokbro
Nano-Science Center, University of Copenhagen, Universitetsparken 5, Bldg. D, DK-2100 Copenhagen, Denmark

(Dated: April 14, 2008)

The Wave Function Matching (WFM) technique has recently been developed for the calculation of electronic transport in quantum two-probe systems. In terms of efficiency it is comparable with the widely used Green's function approach. The WFM formalism presented so far requires the evaluation of all the propagating and evanescent bulk states of the left and right electrodes in order to obtain the correct coupling between device and electrode regions. In this paper we will describe a modified WFM approach that allows for the exclusion of the vast majority of the evanescent states in all parts of the calculation. This approach makes it feasible to apply iterative techniques to efficiently determine the few required bulk states, which allows for a significant reduction of the computational expense of the WFM method. We illustrate the efficiency of the method on a carbon nanotube field-effect-transistor (FET) device displaying band-to-band tunneling and modeled within the semi-empirical Extended Hückel theory (EHT) framework.

PACS numbers: 73.40.-c, 73.63.-b, 72.10.-d, 85.35.Kt, 85.65.+h

I. INTRODUCTION

Quantum transport simulations have become an important theoretical tool for investigating the electrical properties of nano-scale systems.[1–5] The basis for the approach is the Landauer-Büttiker picture of coherent transport, where the electrical properties of a nano-scale constriction is described by the transmission coefficients of a number of one-electron states propagating coherently through the constriction. The approach has been used successfully to describe the electrical properties of a wide range of nano-scale systems, including atomic wires, molecules and interfaces.[6–15] In order to apply the method to semiconductor device simulation, it is necessary to handle systems comprising millions of atoms, and this will require new efficient algorithms for calculating the transmission coefficient.

Our main purpose in this paper is to give details of a method we have developed, based on the WFM technique,[16–18], which is suitable for studying electronic transport in large-scale atomic two-probe systems, such as large carbon nanotubes or nano-wire configurations.

To our knowledge, the WFM schemes presented so far in the literature requires the evaluation of all the Bloch and evanescent bulk modes of the left and right electrodes in order to obtain the correct coupling between device and electrode regions. The reason for this is that it requires the complete set of bulk modes to be able to represent the proper reflected and transmitted wave functions. In this paper we will describe a modified WFM approach that allows for the exclusion of the vast majority of the evanescent modes in all parts of the calculation by simply extending the central region with a few layers. This approach makes it feasible to apply iterative techniques (e.g, as described in[19]) to efficiently determine the relatively few bulk modes of interest, which allows for a significant reduction of the computational expense of the WFM method in practice.

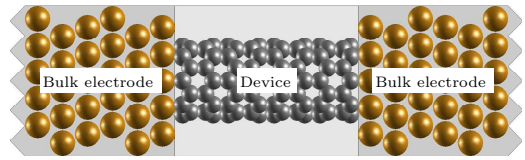


FIG. 1: (Color online) Schematic illustration of a nano-scale two-probe system in which a device is sandwiched between two semi-infinite bulk electrodes.

We adopt the many-channel formulation of Landauer and Büttiker to describe electron transport in nano-scale two-probe systems composed of a left and a right electrode attached to a central device, see Fig. 1. In this formulation, the conduction $\mathcal{G}$ of incident electrons through the device is intuitively given in terms of transmission and reflection matrices, $\boldsymbol{t}$ and $\boldsymbol{r}$, that satisfy the unitarity condition $\boldsymbol{t}^\dagger\boldsymbol{t} + \boldsymbol{r}^\dagger\boldsymbol{r} = \mathbf{1}$ in the case of elastic scattering. The matrix element t_{ij} is the probability amplitude of an incident electron in a state i in the left electrode being scattered into a state j in the right electrode, and correspondingly r_{ik} is the probability of it being reflected

back into state k in the left electrode. This simple interpretation yields the Landauer-Büttiker formula[3]

$$\mathcal{G} = \frac{2e^2}{h}\mathrm{Tr}[\boldsymbol{t}^\dagger \boldsymbol{t}], \tag{1}$$

which holds in the limit of infinitesimal voltage bias and zero temperature.

The WFM method is based upon direct matching of the bulk modes in the left and right electrode to the scattering wave function of the central region. For the most part this involves two major tasks; obtaining the bulk electrode modes and solving a system of linear equations. The available modes in the left and right electrodes are the solutions from the corresponding ideal electrodes. These solutions can be characterized as either propagating or evanescent (exponentially decaying) states but only the propagating states contribute to $\mathcal{G}$ in Eq. (1). We may write $\mathcal{G} = (2e^2/h)T$, where

$$T = \sum_{kk'} |t_{kk'}|^2 \tag{2}$$

is the total transmission and the sum is limited to propagating states k and k' in the left and right electrode, respectively. Notice, however, that the evanescent states are still needed in order to obtain the correct matrix elements $t_{kk'}$. We will discuss this matter in Sect. III C.

The rest of the paper is organized as follows. The WFM formalism used to obtain $\boldsymbol{t}$ and $\boldsymbol{r}$ is introduced in Sect. II. In Sect. III we present our method to effectively exclude the rapidly decaying evanescent states from the two-probe transport calculations. Numerical results are presented in Sect. IV. and the paper ends with a short summary and outlook.

II. FORMALISM

In this section we give a minimal review of the formalism and notation that is used in the current work. To determine the transmission and reflection matrices $\boldsymbol{t}$ and $\boldsymbol{r}$ for our two-probe systems we will apply the recently developed wave function matching (WFM) method.[16–18,20] This technique has several attractive features compared to the more widely used and mathematically equivalent Green's function approach.[1,2] Most importantly, the transparent Landauer picture of electrons scattering via the central region between Bloch states of the electrodes is retained throughout the calculation. Moreover, WFM allows one to consider the significance of each available state individually in order to achieve more efficient numerical procedures to obtain $\boldsymbol{t}$ and $\boldsymbol{r}$.

A. Wave function matching

Let us assume a tight-binding setup for the two-probe systems in which the infinite structure is divided into

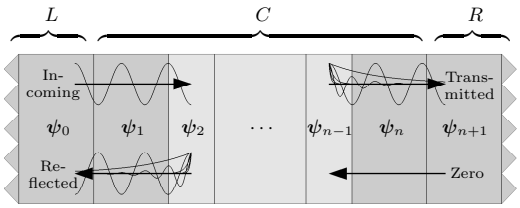


FIG. 2: (Color online) Schematic representation of WFM applied to layered two-probe systems, where the central device region, consisting of layers $i = 1, \ldots, n$, is attached to left and right semi-infinite electrodes. The incoming propagating state from the left electrode is scattered in the central region and end up as reflected and transmitted superpositions of propagating and evanescent states.

principle layers numbered $i = -\infty, \ldots, \infty$ and composed of a finite central (C) region containing the device and two semi-infinite left (L) and right (R) electrode regions, see Fig. 2. The wave function is $\boldsymbol{\psi}_i(\boldsymbol{x}) = \sum_j^{m_i} c_{i,j}\chi_{i,j}(\boldsymbol{x} - \boldsymbol{X}_{i,j})$ in layer i, where $\chi_{i,j}$ denotes localized non-orthogonal atomic orbitals and $\boldsymbol{X}_{i,j}$ are the positions of the m_i atoms. We represent $\boldsymbol{\psi}_i(\boldsymbol{x})$ by a column vector of the expansion coefficients, given by $\boldsymbol{\psi}_i = [c_{i,1}, \ldots, c_{i,m_i}]^T$, and write the wave function $\boldsymbol{\psi}$ extending over the entire system as $\boldsymbol{\psi} = [\boldsymbol{\psi}_{-\infty}^{\mathrm{T}}, \ldots, \boldsymbol{\psi}_{\infty}^{\mathrm{T}}]^{\mathrm{T}}$. We also assume that the border layers 1 and n of the central region are always identical to a layer of the connecting electrodes.

We refer the reader to Refs. 16–18 for details on how to employ WFM to our setup. Here and in the rest of this paper, we will use the following notation for the key elements: The matrices $\boldsymbol{\Phi}_L^{\pm} = [\boldsymbol{\phi}_{L,1}^{\pm}, \ldots, \boldsymbol{\phi}_{L,m_L}^{\pm}]$ contain in their columns the full set of m_L left-going ($-$) and m_L right-going ($+$) bulk states $\boldsymbol{\phi}_{L,k}^{\pm}$ of the left electrode, and the diagonal matrices $\boldsymbol{\Lambda}_L^{\pm} = \mathrm{diag}[\lambda_{L,1}^{\pm}, \lambda_{L,2}^{\pm}, \ldots, \lambda_{L,m_L}^{\pm}]$ hold the corresponding Bloch factors.[29] If trivial states with $|\boldsymbol{\phi}_{L,k}^{+}| = \boldsymbol{0}$ or $|\boldsymbol{\phi}_{L,k}^{-}| = \texttt{inf}$ occur in practice these are simply rejected. We assume that all the evanescent bulk states are (state-)normalized $\boldsymbol{\phi}_{L,k}^{\pm\dagger}\boldsymbol{\phi}_{L,k}^{\pm} = 1$, while all the Bloch bulk states are flux-normalized[30] $\boldsymbol{\phi}_{L,k}^{\pm\dagger}\boldsymbol{\phi}_{L,k}^{\pm} = d_L/v_{L,k}^{\pm}$, where $v_{L,k}^{\pm}$ are the group velocities[15,21] and d_L is the layer thickness. Similarly for the right electrode the matrices $\boldsymbol{\Phi}_R^{\pm}$ and $\boldsymbol{\Lambda}_R^{\pm}$ are formed. We can then define the Bloch matrices[17] as $\boldsymbol{B}_L^{\pm} = \boldsymbol{\Phi}_L^{\pm}\boldsymbol{\Lambda}_L^{\pm}(\boldsymbol{\Phi}_L^{\pm})^{-1}$ and $\boldsymbol{B}_R^{\pm} = \boldsymbol{\Phi}_R^{\pm}\boldsymbol{\Lambda}_R^{\pm}(\boldsymbol{\Phi}_R^{\pm})^{-1}$. The system of linear equations for $\boldsymbol{\psi}_C$ is subsequently written as

$$(E\boldsymbol{S}_C - \boldsymbol{H}_C)\boldsymbol{\psi}_C = \boldsymbol{b}, \tag{3}$$

where E is the energy, $\boldsymbol{\psi}_C = [\boldsymbol{\psi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\psi}_n^{\mathrm{T}}]^{\mathrm{T}}$ is the central

region wave function, and $\boldsymbol{S}_C$ and

$$\boldsymbol{H}_C = \begin{pmatrix} \boldsymbol{H}_1 + \boldsymbol{\Sigma}_L & \boldsymbol{H}_{1,2} & & & \\ \boldsymbol{H}_{1,2}^\dagger & \boldsymbol{H}_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \boldsymbol{H}_{n-1} & \boldsymbol{H}_{n-1,n} \\ & & & \boldsymbol{H}_{n-1,n}^\dagger & \boldsymbol{H}_n + \boldsymbol{\Sigma}_R \end{pmatrix} \tag{4}$$

are the tight-binding overlap and Hamiltonian matrices of the central region. The right-hand side source term $\boldsymbol{b} = [\boldsymbol{b}_1^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}}, \ldots, \boldsymbol{0}^{\mathrm{T}}]^{\mathrm{T}}$ is specified by the expression

$$\boldsymbol{b}_1 = -(\bar{\boldsymbol{H}}_{0,1}^\dagger + \boldsymbol{\Sigma}_L \boldsymbol{B}_L^+)\boldsymbol{\psi}_0, \tag{5}$$

where $\boldsymbol{\psi}_0$ is the incoming wave function (notice that this source term is defined for layer 1 and not layer 0, as is the case in Refs. 18 and 17). In Eq. (5) we have introduced the overline notation $\bar{\boldsymbol{H}}_i \equiv E\boldsymbol{S}_i - \boldsymbol{H}_i$ and $\bar{\boldsymbol{H}}_{i,j} \equiv E\boldsymbol{S}_{i,j} - \boldsymbol{H}_{i,j}$ (also used below) to enhance the readability.

The matrices $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_R$ are the left and right self-energy matrices. We stress that for the current setup, these matrices are *identical* to the self-energy matrices introduced in the Green's function formalism[1] (to within an infinitesimal imaginary shift of E), and may be evaluated by well-known recursive techniques[22,23] or, more conveniently for WFM, in terms of the Bloch matrices:[16,17]

$$\boldsymbol{\Sigma}_L = \bar{\boldsymbol{H}}_{0,1}^\dagger (\bar{\boldsymbol{H}}_1 + \bar{\boldsymbol{H}}_{0,1}^\dagger (\boldsymbol{B}_L^-)^{-1})^{-1} \bar{\boldsymbol{H}}_{0,1}, \tag{6}$$

and

$$\boldsymbol{\Sigma}_R = \bar{\boldsymbol{H}}_{n,n+1} (\bar{\boldsymbol{H}}_n + \bar{\boldsymbol{H}}_{n,n+1} \boldsymbol{B}_R^+)^{-1} \bar{\boldsymbol{H}}_{n,n+1}^\dagger, \tag{7}$$

For notational simplicity in the following sections, we leave out the implied subscripts L or R, indicating the left or right electrode, whenever the formalism is the same for both (e.g, for symbols $m, \lambda_k, \boldsymbol{\phi}_k, \boldsymbol{\Phi}^\pm, \boldsymbol{\Lambda}^\pm, \boldsymbol{B}^\pm, \boldsymbol{\Sigma}$, etc.).

B. Transmission and reflection coefficients

As a final step we want to determine the $\boldsymbol{t}$ and $\boldsymbol{r}$ matrices from the boundary wave functions $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_n$ that have been obtained by solving Eq. (3).

When the incoming wave $\boldsymbol{\psi}_0$ is specified to be the kth right-going state $\boldsymbol{\phi}_{L,k}^+$ of the left electrode, we can evaluate the kth column of the transmission matrix $\boldsymbol{t}_k$ by solving

$$\boldsymbol{\Phi}_R^+ \boldsymbol{t}_k = \boldsymbol{\psi}_n, \tag{8}$$

where $\boldsymbol{\Phi}_R^+$ is the $m_R \times m_R$ column matrix holding the right-going bulk states of the right electrode (and here assumed to be non-singular). Similarly the kth column of the reflection matrix $\boldsymbol{r}_k$ is given by

$$\boldsymbol{\Phi}_L^- \boldsymbol{r}_k = \boldsymbol{\psi}_1 - \lambda_{L,k}^+ \boldsymbol{\phi}_{L,k}^+, \tag{9}$$

where $\boldsymbol{\Phi}_L^-$ holds the left-going bulk states of the left electrode. The flux normalization ensures that $\boldsymbol{t}^\dagger \boldsymbol{t} + \boldsymbol{r}^\dagger \boldsymbol{r} = \boldsymbol{1}$.[1]

TABLE I: CPU times in seconds when using WFM for calculating $\boldsymbol{t}$ and $\boldsymbol{r}$ at 20 different energies inside $E \in [-2\,\mathrm{eV}; 2\,\mathrm{eV}]$ for various two-probe systems. The numbers of atoms in the central region (electrode unit cell) are indicated. The two right-most columns show the percentage of the CPU time used for computing the electrode bulk states with DGEEV vs. solving the central region linear systems in Eq. (3).

System	Atoms	CPU	DGEEV	Eq. (3)
Li–Li	32(8)	0.1	92%	7%
Li–C chain	91(54|2)	2.8	91%	8%
Al–C×7–Al	74(18)	6.5	87%	11%
Au–BDT–Au	102(27)	171.3	90%	8%
Au–CNT(8,0)×5–Au	268(27)	241.8	65%	34%
CNT(8,0)–CNT(8,0)	192(64)	257.9	95%	4%
CNT(5,0)–CNT(10,0)	300(40|80)	255.1	79%	18%
CNT(18,0)–CNT(18,0)	576(144)	994.1	90%	8%

III. EXCLUDING EVANESCENT STATES

The most time consuming task of the WFM method is to determine the electrode states, which requires solving a quadratic eigenvalue problem.[16] As examples, see the profiling results listed in Table I, where we have used the method to compute $\boldsymbol{t}$ and $\boldsymbol{r}$ for a selection of two-probe systems.[31] The CPU timings show that to determine the electrode states by employing the state-of-the-art LAPACK eigensolver DGEEV is, in general, much more expensive than to solve the system of linear equations in Eq. (3). We expect this trend to hold for larger systems as well. Therefore, in the attempt to model significantly larger devices (thousands of atoms), it is of essential interest to reduce the numerical cost of the electrode states calculation. We argue that a physically reasonable approach is to limit the number of electrode states taken into account, e.g., by excluding the least important evanescent states. In this section, a new technique to do this in a rigorous and systematic fashion is presented.

A. Decay of evanescent states

The procedure to determine the Bloch factors λ_k and non-trivial states $\boldsymbol{\phi}_k$ of an ideal electrode and subsequently characterize these as right-going (+) or left-going (−) is well described in the literature.[16–18,24] We note that only the obtained propagating states with $|\lambda_k| = 1$ are able to carry charge deeply into the electrodes and thus enter the Landauer expression in Eq. (2). The evanescent states with $|\lambda_k| \neq 1$, on the other hand, decay exponentially but can still contribute to the current in a two-probe system, as the "tails" may reach across the central region boundaries.

Consider a typical example of an electrode states evaluation: We look at a gold electrode with 27 atoms in the unit cell represented by 9 ($\mathrm{sp^3d^5}$) orbitals for each Au-

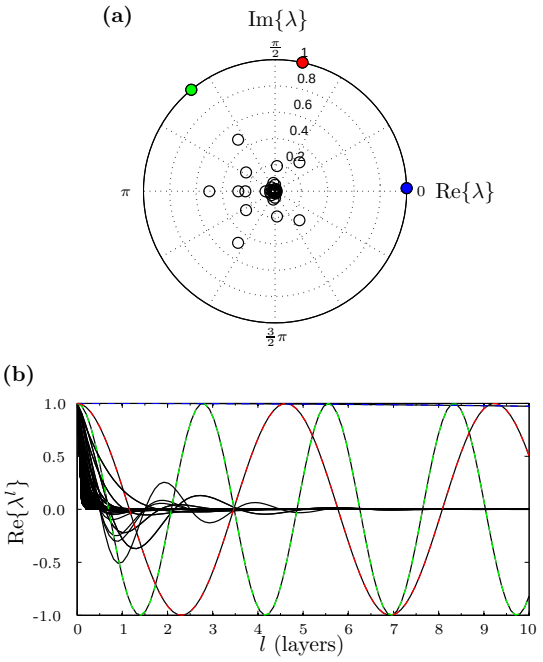


FIG. 3: (Color online) (a) Positions of the Bloch factors λ_k ($|\lambda_k| \leq 1$) obtained for a bulk Au(111) electrode with 27 atoms per unit cell at $E = -1.5$ eV. (b) Amplitudes of the corresponding normalized electrode states $\boldsymbol{\phi}_k$ moving through 10 layers of the ideal bulk electrode. A total of 243 states are shown of which 3 are propagating (colored/dashed) and the rest are evanescent (circles/black).

atom. Such a system results in 243 right-going and 243 left-going states. Fig. 3a shows the positions in the complex plane of the Bloch factors corresponding the right-going states (i.e., $|\lambda_k| \leq 1$) for energy $E = -1.5$ eV. We see that there are exactly three propagating states, which have Bloch factors located on the unit circle. The remaining states are evanescent, of which many have Bloch factors with small magnitude very close to the origin.

Fig. 3b illustrates how the 243 left-going states would propagate through 10 successive gold electrode unit cells. The figure shows that the amplitudes of the three propagating states are unchanged, while the evanescent states are decaying exponentially. In particular, we note that the evanescent states with Bloch factors of small magnitude are very rapidly decaying and vanishes in comparison to the propagating states after only a few layers. In the following, we will exploit this observation and attempt to exclude such evanescent states from the WFM calculation altogether. Formally this can be accomplished if only the electrode states $\boldsymbol{\phi}_k$ with Bloch factors λ_k satisfying

$$\lambda_{\min} \leq |\lambda_k| \leq \lambda_{\min}^{-1}, \tag{10}$$

are computed and subsequently taken into account, for a reasonable choice of $0 < \lambda_{\min} < 1$. Eq. (10) is therefore adopted in the coming sections as the key relation to select a particular subset of the available electrode states.

B. Extra electrode layers

We will denote the state matrices from which the rapidly decaying evanescent states are excluded via Eq. (10), and also the Bloch matrices and self-energy matrices obtained from these, with a tilde, i.e., as $\tilde{\boldsymbol{\Phi}}^{\pm}$, $\tilde{\boldsymbol{B}}^{\pm}$ and $\tilde{\boldsymbol{\Sigma}}$. The state matrices holding the excluded states are denoted by a math-ring accent $\mathring{\boldsymbol{\Phi}}^{\pm}$, so that

$$\boldsymbol{\Phi}^{\pm} = [\tilde{\boldsymbol{\Phi}}^{\pm}, \mathring{\boldsymbol{\Phi}}^{\pm}], \tag{11}$$

is the assumed splitting of the full set. All expressions to evaluate the Bloch and self-energy matrices are unchanged as given in Sect. II (now $(\tilde{\boldsymbol{\Phi}}^{\pm})^{-1}$ merely represents the *pseudo-inverses* of $\tilde{\boldsymbol{\Phi}}^{\pm}$). However, since the column spaces of $\tilde{\boldsymbol{\Phi}}^{\pm}$ are not complete, there is no longer any guaranty that WFM can be performed so that the resulting self-energy matrices and, in turn, the solution $\boldsymbol{\psi}_C = [\boldsymbol{\psi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\psi}_n^{\mathrm{T}}]^{\mathrm{T}}$ of the linear system in Eq. (3), are correct. In addition, it is clear that errors can occur in the calculation of $\boldsymbol{t}$ and $\boldsymbol{r}$ from Eqs. (8) and (9) because the boundary wave functions $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_n$ might not be fully represented in the reduced sets $\tilde{\boldsymbol{\Phi}}_R^{+}$ and $\tilde{\boldsymbol{\Phi}}_L^{-}$.

As explicitly shown in Refs. 16–18, the key to deriving Eq. (3) is twofold: fixing the layer wave functions coming into the C region (e.g., in our case $\boldsymbol{\psi}_1^{+} = \lambda_{L,k}^{+}\boldsymbol{\phi}_{L,k}^{+}$ and $\boldsymbol{\psi}_n^{-} = \boldsymbol{0}$) and matching the layer wave functions across the C region boundaries (we remind the reader that our setup has one more electrode layer on both sides of C compared to the setup of Refs. 18 and 17).

The matching is accomplished by using the $\boldsymbol{B}^{\pm}$ matrices, which by construction propagate the layer wave functions in the bulk electrode,[16–18] i.e.,

$$\boldsymbol{\psi}_j^{\pm} = (\boldsymbol{B}^{\pm})^{j-i}\boldsymbol{\psi}_i^{\pm}, \tag{12}$$

where subscript L is implied for the left electrode ($i, j \leq 1$), and R for the right electrode ($i, j \geq n$). Notice that the Bloch matrices are always square and also invertible since any trivial electrode are rejected from the outset in the current formalism. When the reduced Bloch matrices $\tilde{\boldsymbol{B}}^{\pm}$ are used instead of $\boldsymbol{B}^{\pm}$, however, the possible components of the wave functions outside the column spaces of $\tilde{\boldsymbol{\Phi}}^{\pm}$ are not properly matched, and the boundary conditions are not necessarily satisfiable.

In order to diminish the errors introduced by excluding evanescent states, we propose to insert additional electrode layers in the central region, see Fig. 4. As illustrated in the previous section, this would quickly reduce

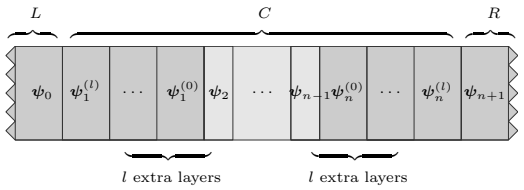


FIG. 4: (Color online) Two-probe system in which the C region boundaries are expanded by l extra electrode layers.

the imprint of the rapidly decaying evanescent states in the boundary layer wave functions $\tilde{\boldsymbol{\psi}}_1$ and $\tilde{\boldsymbol{\psi}}_n$, which means that the critical components outside the column spaces $\tilde{\boldsymbol{\Phi}}^{\pm}$ becomes negligible at an exponential rate in terms of the number of additional layers. We emphasize that the inserted layers may be "fictitious" in the sense that they can be accommodated by simple block-Gaussian-eliminations prior to the solving of Eq. (3) for the original system.

The above statements are confirmed by the following analysis. In the particular case, where l extra electrode layers are inserted and the border layers of the C region are identical to the connecting electrode layers, we can write the boundary matching equations as[16–18]

$$\boldsymbol{\psi}_0 = (\tilde{\boldsymbol{B}}_L^+)^{-1}\boldsymbol{\psi}_1^{(l)+} + (\tilde{\boldsymbol{B}}_L^-)^{-1}\boldsymbol{\psi}_1^{(l)-} \tag{13}$$

for the left boundary and

$$\boldsymbol{\psi}_{n+1} = \tilde{\boldsymbol{B}}_R^+\boldsymbol{\psi}_n^{(l)+} + \tilde{\boldsymbol{B}}_R^-\boldsymbol{\psi}_n^{(l)-} \tag{14}$$

for the right boundary, where $\boldsymbol{\psi}_1^{(l)+} = \lambda_{L,k}^+\boldsymbol{\phi}_{L,k}^+$ and $\boldsymbol{\psi}_n^{(l)-} = \mathbf{0}$ are fixed as boundary conditions. We point out, that the l extra layers are bulk layers extending from each electrode and therefore connected via the relation in Eq. (12) for L and R, respectively. Moreover, since the electrode wave functions can always be expanded in the corresponding complete set of bulk states, we may write

$$\boldsymbol{\psi}_i^{\pm} = \boldsymbol{\Phi}^{\pm}\boldsymbol{a}_i^{\pm} = [\tilde{\boldsymbol{\Phi}}^{\pm}, \mathring{\boldsymbol{\Phi}}^{\pm}]\begin{bmatrix}\tilde{\boldsymbol{a}}_i^{\pm}\\ \mathring{\boldsymbol{a}}_i^{\pm}\end{bmatrix}, \tag{15}$$

where $\boldsymbol{a}_i^{\pm} = [\tilde{\boldsymbol{a}}_i^{\pm T}, \mathring{\boldsymbol{a}}_i^{\pm T}]^T$ are vectors that contain the expansion coefficients and subscript L is implied for the left electrode ($i \leq 1$), and R for the right electrode ($i \geq n$). Thus we may consider the (unfixed) boundary wave functions entering Eqs. (13) and (14), by explicitly writing

$$\boldsymbol{\psi}_1^{(l)-} = (\boldsymbol{B}_L^-)^{-l}\boldsymbol{\psi}_1^- = [\tilde{\boldsymbol{\Phi}}_L^-, \mathring{\boldsymbol{\Phi}}_L^-]\begin{bmatrix}(\tilde{\boldsymbol{\Lambda}}_L^-)^{-l}\tilde{\boldsymbol{a}}_1^-\\ (\mathring{\boldsymbol{\Lambda}}_L^-)^{-l}\mathring{\boldsymbol{a}}_1^-\end{bmatrix}, \tag{16}$$

and

$$\boldsymbol{\psi}_n^{(l)+} = (\boldsymbol{B}_R^+)^{l}\boldsymbol{\psi}_n^+ = [\tilde{\boldsymbol{\Phi}}_R^+, \mathring{\boldsymbol{\Phi}}_R^+]\begin{bmatrix}(\tilde{\boldsymbol{\Lambda}}_R^+)^{l}\tilde{\boldsymbol{a}}_n^+\\ (\mathring{\boldsymbol{\Lambda}}_R^+)^{l}\mathring{\boldsymbol{a}}_n^+\end{bmatrix}, \tag{17}$$

using the definition $\boldsymbol{B}^{\pm} = \boldsymbol{\Phi}^{\pm}\boldsymbol{\Lambda}^{\pm}(\boldsymbol{\Phi}^{\pm})^{-1}$. This shows that the critical components outside the column spaces of $\tilde{\boldsymbol{\Phi}}_L^{\pm}$ and $\tilde{\boldsymbol{\Phi}}_R^{\pm}$ are given by coefficients $(\mathring{\boldsymbol{\Lambda}}_L^-)^{-l}\mathring{\boldsymbol{a}}_1^-$ and $(\mathring{\boldsymbol{\Lambda}}_R^+)^{l}\mathring{\boldsymbol{a}}_n^+$, respectively, and assuming we exclude fastest decaying of the evanescent states according to Eq. (10), that is, $|\lambda_k| > \lambda_{\min}^{-1}$ for the diagonal elements of $\mathring{\boldsymbol{\Lambda}}_L^-$ and $|\lambda_k| < \lambda_{\min}$ for the diagonal elements of $\mathring{\boldsymbol{\Lambda}}_R^+$, where $\lambda_{\min}$ is less than 1, these coefficients always decrease as a function of l.

We conclude that WFM with the reduced Bloch matrices $\tilde{\boldsymbol{B}}^{\pm}$ approaches the exact case with $\boldsymbol{B}^{\pm}$ if additional electrode layers are inserted as suggested, and therefore, that the solution $\tilde{\boldsymbol{\psi}}_C$ obtained from Eq. (3) when only a reduced set of bulk states are used, approaches the correct solution $\boldsymbol{\psi}_C$ accordingly.

C. Accuracy

Since it is clear that the exclusion of some of the evanescent states may introduce errors when using the WFM method, it is important to be able to estimate and monitor the accuracy of the results obtained. We now discuss how this can be done in a systematic fashion in terms of the parameter $\lambda_{\min}$ and the number l of extra electrode layers.

Let us first consider the accuracy of the transmission matrix $\boldsymbol{t}$. Suppose that the rapidly decaying evanescent states $\mathring{\boldsymbol{\Phi}}^{\pm}$ are excluded in the WFM as demonstrated in Sect. III B, but still available to evaluate the transmission coefficients with Eq. (8), which here becomes $\boldsymbol{t}_k = [\tilde{\boldsymbol{\Phi}}_R^+, \mathring{\boldsymbol{\Phi}}_R^+]^{-1}\boldsymbol{\psi}_n^{(l)+}$, since $\boldsymbol{\psi}_n^{(l)-} = \mathbf{0}$ from the boundary conditions. Inserting Eq. (17) in this expression, we see that the first $\tilde{m}_R$ elements of the column vector $\boldsymbol{t}_k$ are given by $\tilde{\boldsymbol{t}}_k = (\tilde{\boldsymbol{\Lambda}}_R^+)^{l}\tilde{\boldsymbol{a}}_n^+$. These coefficients are exact, assuming that enough extra layers are inserted to accomodate accurate WFM for the system.

We now wish to compare the exact coefficients with the ones obtained from $\tilde{\boldsymbol{t}}_k' = (\tilde{\boldsymbol{\Phi}}_R^+)^{-1}\boldsymbol{\psi}_n^{(l)+}$, that is, when the rapidly decaying evanescent states are excluded from the calculations altogether. In order to do this, we use the property of the pseudo-inverse which allows us to write

$$(\tilde{\boldsymbol{\Phi}}_R^+)^{-1}[\tilde{\boldsymbol{\Phi}}_R^+, \mathring{\boldsymbol{\Phi}}_R^+] = [\tilde{\boldsymbol{I}}, (\tilde{\boldsymbol{\Phi}}_R^+)^{-1}\mathring{\boldsymbol{\Phi}}_R^+], \tag{18}$$

where $\tilde{\boldsymbol{I}}$ is the identity matrix of order equal to the number of included states $\tilde{m}_R$. From the expression in Eq. (17) it then follows that

$$\tilde{\boldsymbol{t}}_k' = \tilde{\boldsymbol{t}}_k + (\tilde{\boldsymbol{\Phi}}_R^+)^{-1}\mathring{\boldsymbol{\Phi}}_R^+(\mathring{\boldsymbol{\Lambda}}_R^+)^{l}\mathring{\boldsymbol{a}}_n^+, \tag{19}$$

which corresponds to the correct coefficients $\tilde{\boldsymbol{t}}_k$ plus an error term.

We have already established in the previous section that the $(\mathring{\boldsymbol{\Lambda}}_R^+)^{l}\mathring{\boldsymbol{a}}_n^+$ factor in the error term will decrease as a function of l. To ascertain that the total error term also decreases we look at the 2-norm of $(\tilde{\boldsymbol{\Phi}}_R^+)^{-1}\mathring{\boldsymbol{\Phi}}_R^+$, which will satisfy $||(\tilde{\boldsymbol{\Phi}}_R^+)^{-1}\mathring{\boldsymbol{\Phi}}_R^+||_2 \leq \mathring{m}_R^{\frac{1}{2}}||(\tilde{\boldsymbol{\Phi}}_R^+)^{-1}||_2$, since

$||\mathring{\mathbf{\Phi}}_R^+||_2 \leq \mathring{m}_R^{\frac{1}{2}}$ when all evanescent states are assumed normalized. The norm $||(\tilde{\mathbf{\Phi}}_R^+)^{-1}||_2$ can be readily evaluated and depends on the set of states included via the parameter $\lambda_{\min}$ but not on l. We then have that $(\tilde{\mathbf{\Phi}}_R^+)^{-1}\mathring{\mathbf{\Phi}}_R^+$ is independent of l.

Writing Eq. (19) as $\tilde{\boldsymbol{t}}'_k = \tilde{\boldsymbol{t}}_k + \tilde{\boldsymbol{\epsilon}}_k$, where $\tilde{\boldsymbol{\epsilon}}_k$ holds the errors on the coefficients of the kth column, we further obtain that the corresponding total transmission T' obtained from Eq. (2) can be expressed as

$$T' = T + \sum_{kk'} (\tilde{t}^*_{kk'}\tilde{\epsilon}_{kk'} + \tilde{\epsilon}^*_{kk'}\tilde{t}_{kk'} + |\tilde{\epsilon}_{kk'}|^2) \tag{20}$$

where T is the exact result and the summation is over the Bloch states k and k' in the left and right electrode, respectively.

In the attempt to estimate the order of the error term in Eq. (20) we may (as a worst case approximation) take all diagonal elements of $\mathring{\mathbf{\Lambda}}_R^+$ to be equal to the maximum range $\lambda_{\min}$ of Eq. (10), which makes all elements $\tilde{\epsilon}_{kk'}$ proportional to $\lambda^l_{\min}$. Thus we arrive at the simple relation

$$|T' - T| \sim \lambda^l_{\min} + \mathcal{O}\big((\lambda^l_{\min})^2\big), \tag{21}$$

which shows that the error decreases exponentially in terms of the number of extra layers l.

In practice, Eq. (21) can be adopted as a reasonable order of magnitude estimate of the accuracy of T'. Alternatively, we are able to directly monitor the error arising on the boundary conditions, e.g., in terms of the coefficient vectors $\tilde{\boldsymbol{b}}_{L,k} \equiv (\tilde{\mathbf{\Phi}}_R^+)^{-1}(\boldsymbol{\psi}_1^{(l)+} - \lambda^+_{L,k}\boldsymbol{\phi}^+_{L,k})$ and $\tilde{\boldsymbol{b}}_{R,k} \equiv (\tilde{\mathbf{\Phi}}_R^-)^{-1}\boldsymbol{\psi}_n^{(l)-}$, where $\boldsymbol{\psi}_1^{(l)+}$ and $\boldsymbol{\psi}_n^{(l)-}$ are given by solving Eq. (3). It is clear that $|\tilde{\boldsymbol{b}}_{L,k}| = 0$ and $|\tilde{\boldsymbol{b}}_{R,k}| = 0$ in the case where the boundary conditions are exactly satisfied. Taking into account the similarity in the expressions for $\tilde{\boldsymbol{t}}'_k$ and $\tilde{\boldsymbol{b}}_{R,k}$ and assuming a similar order of errors in $\boldsymbol{\psi}_n^{(l)+}$ and $\boldsymbol{\psi}_n^{(l)-}$, we would also expect the same order of magnitude of $|\tilde{\boldsymbol{\epsilon}}_k|$ and $|\tilde{\boldsymbol{b}}_{R,k}|$. This suggests another order of magnitude accuracy estimate from Eq. (20), which is strait forward to monitor using the results available with the reduced set of bulk states:

$$|T'-T| \leq \sum_k (2|\tilde{\boldsymbol{t}}_k||\tilde{\boldsymbol{\epsilon}}_k| + |\tilde{\boldsymbol{\epsilon}}_k|^2) \sim \sum_k (2|\tilde{\boldsymbol{t}}_k||\tilde{\boldsymbol{b}}_{R,k}| + |\tilde{\boldsymbol{b}}_{R,k}|^2), \tag{22}$$

where all the vector norms (e.g., $|\tilde{\boldsymbol{t}}_k|^2 = \sum_{k'} |\tilde{t}_{kk'}|^2$) are assumed to be taken over the elements corresponding to propagating bulk states k' only.

Finally, we note without explicit derivation, that similar arguments for the reflection matrix with columns $\tilde{\boldsymbol{r}}'_k = (\tilde{\mathbf{\Phi}}_L^-)^{-1}(\boldsymbol{\psi}_1^{(l)-} - \lambda^+_{L,k}\boldsymbol{\phi}^+_{L,k})$ and the total reflection coefficient R', as presented above for $\tilde{\boldsymbol{t}}'_k$ and T', results in the same accuracy expressions for $|R' - R|$ as for $|T' - T|$ in Eqs. (21) and (22), if we substitute $\tilde{\boldsymbol{t}}_k \to \tilde{\boldsymbol{r}}_k$ and $\tilde{\boldsymbol{b}}_{R,k} \to \tilde{\boldsymbol{b}}_{L,k}$.

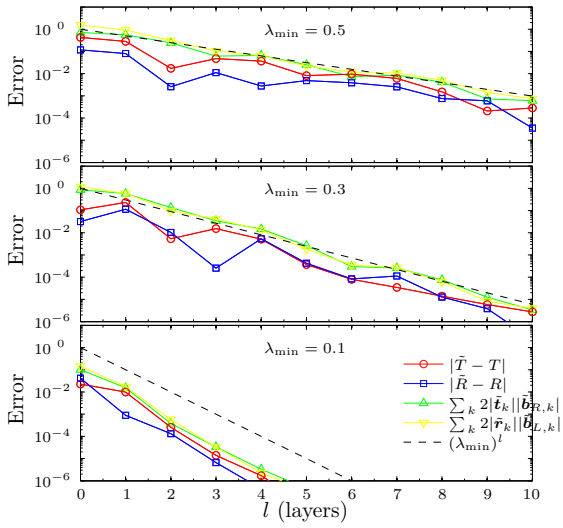


FIG. 5: (Color online) Error (absolute) in the calculated total transmission and reflection coefficients T' and R' as a function of l. The panels show the cases of $\lambda_{\min}$ set to 0.5, 0.3 and 0.1, which corresponds to 3, 14 and 31 Au bulk states (out of 243, see Fig. 3) taken into account, respectively. The dashed line indicate the theoretical accuracy estimate $\lambda^l_{\min}$.

D. Example

To end this section, we exemplify the previous discussion quantitatively by looking at the Au(111) electrode described earlier, and assuming a 128 atom (4 unit cells) device of zigzag-(8,0) carbon nano tube (CNT), see the configuration in Fig. 1. For energy $E = -1.5$ eV, we have calculated the deviation between the total transmission obtained when all bulk states are taken into account (T) and when some evanescent states are excluded (T') as specified with different settings of $\lambda_{\min}$. Deviations are also determined for the corresponding total reflection coefficients (R and R'). Fig. 5 shows the results as a function of l, together with the estimate $\lambda^l_{\min}$ of Eq. (21) and the estimate of Eq. (22) both for the transmission and reflection coefficients, where the higher order terms have been neglected,

We observe that the absolute error in the obtained transmission coefficients (red curves) and reflection coefficients (blue curves) are generally decreasing as a function of l, following the same convergence rate as $\lambda^l_{\min}$ (dashed line). Looking closer at results for neighbor l values, we see that the errors initially exhibit wave-like oscillations. This is directly related to the wave form of the evanescent states that have been excluded (see the propagation of the slowest decaying black curves in Fig. 3(b)), since the representation of these states in the

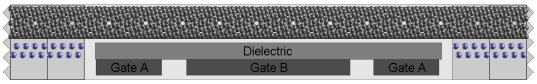


FIG. 6: (Color online) Schematic illustration of a carbon nanotube (8,4) band-to-band tunneling device. The carbon nanotube is positioned on Li surfaces next to an arrangement of three gates.

reduced spaces $\tilde{\mathbf{\Phi}}^{\pm}$ (i.e., the expansion coefficients in $\tilde{\boldsymbol{\epsilon}}_k$) may shift when l is increased. In other words, although the norm of the errors $|\tilde{\boldsymbol{\epsilon}}_k|$ are decreasing as a function of l, the specific error $\tilde{\epsilon}_{kk'}$ on a given (large) coefficient of $\tilde{t}'_{kk'}$ or $\tilde{r}'_{kk'}$ may increase, which means that the overall error term in Eq. (20) can go up. Fortunately, however, this is only a local phenomenon with the global trend being rapidly decreasing errors.

Consider also the quality of the simple accuracy estimate of $\lambda^l_{\min}$ and the estimates expressed by Eq. (22) for the transmission coefficients (green curves) and reflection coefficients (yellow curves), respectively. For relatively large $\lambda_{\min}$ all estimates are very good. However, for smaller values of $\lambda_{\min}$, only the latter two retain a high quality while the $\lambda^l_{\min}$ estimate tends to be overly pessimistic. It is important to remember, that these estimates are by no means strict conditions but very reasonable to make an order of magnitude estimate of the accuracy.

We note in passing, that the results in the top panel of Fig. 5 corresponds to using *only* the propagating Bloch states in the transmission calculation. Still we are able to compute T and R to an absolute accuracy of three digits by inserting 2×5 extra electrode layers in the two-probe system. This is quite remarkable and shows promise for large-scale systems, e.g., with nano-wire electrodes, for which the total number of evanescent states available becomes exceedingly great.

IV. APPLICATION

In this section we will apply the developed method to a nano-device consisting of a CNT stretched between to two metal electrodes and controlled by three gates. The setup is inspired by Appenzeller *et al.*[25], and we expect this particular arrangement to be able to display so-called band-to-band (BTB) tunneling, where one observes gate induced tunneling from the valence band into the conduction band of a semi-conducting CNT and vice versa.

We show the configuration of the band-to-band tunneling two-probe system in Fig. 6. The device configuration contains 10 principal layers of a CNT(8,4), having 112 atoms in each. The diameter of the tube and layer thickness are 8.3 Å and 11.3 Å, respectively. The electrodes consist of CNT(8,4) resting on a thin surfaces of Li, where the lattice constant of the Li layers is stretched to fit the

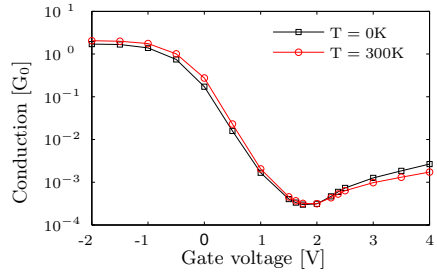


FIG. 7: (Color online) Conduction as a function of the Gate-A voltage in units of the conductance quantum G_0. In the calculations we use a dielectric constant of 4, $V_{\text{Gate-A}} = -2.0$ V, and vary $V_{\text{Gate-B}}$ from -2.0 V to 4.0 V as indicated.

layer thickness of the CNT. The central region of the two-probe system comprises a total of 1440 atoms. An arrangement of rectangular gates are positioned below the carbon nanotube as indicated on the figure. In the plane of the illustration (length × height) the dimensions are as follows: Dielectric 108 Å×5 Å; Gate-A 108 Å×5 Å; Gate-B 20 Å × 5 Å. All the regions are centered with respect to the electrodes so that the complete setup has mirror symmetry. In the direction perpendicular to the illustration the configuration is assumed repeated every 19.5 Å as a super-cell.

We have obtained the electronic density of the BTB device by combining the NEGF formalism with a semi-empirical Extended Hückel theory model (EHT) using the parameterization of Hoffmann.[26] In order to adjust the charge transfer between the CNT and the Li electrodes, such that the Fermi level is just below the conduction band of CNT(8,4) we add the term $\delta\epsilon$S to the Li parameters. This corresponds to an n-type doping. At self-consistency, the average charge transfer from Li to the nanotube is 0.002 e per carbon atom in the electrode and the Fermi energy is located at -4.29 eV, which is only 0.07 eV below the conduction band of CNT(8,4). The electrostatic treatment of the dielectric and gates as part of the self-consistent procedure for obtaining the electronic density is described in our recent publication Stokbro *et al.*.[27]

In the following we show results from a calculation of the transmission spectrum $T(E)$ for $V_{\text{Gate-A}} = -2.0$ V and a dielectric constant of 4. To begin with we calculate the electronic conductance for different Gate-B voltages in the range $[-2\text{ V}, 4\text{ V}]$. The results for temperature T = 0 K, in terms of the unit conduction G_0 are displayed with the black curve in Fig. 7. It shows an initial conductance for $V_{\text{Gate-B}} = -2.0$ V of the order of one, a subsequent drop by four orders of magnitude around $V_{\text{Gate-B}} = 2.0$ V, and a final increase of one order of magnitude towards $V_{\text{Gate-B}} = 4.0$ V. In addition to the zero temperature conduction which is equal to $T(E_F)$,

where E_F is the Fermi energy, we also display the results at room temperature T = 0 K (red curve), which can be obtained from linear response as

$$\mathrm{G} = \int \mathrm{d}E\; T(E) \frac{e^{(E-E_F)/\mathrm{k}_B\mathrm{T}}}{(1+e^{(E-E_F)/\mathrm{k}_B\mathrm{T}})^2} \qquad (23)$$

The overall trend of the conduction curve is similar for room temperature, and can be explained as band-to-band tunneling which is tuned by the gate potentials.

In order for BTB tunneling to appear in CNFETs, fields along the length of the tube have to be created that are strong enough to shift the conductance or valence bands by at least the gap energy of the CNT. In the case of CNT(8,4) the band gap is ~ 0.8 eV which can be transcended via the three-gate arrangement. More specifically, we present in the left part of Fig. 8 the total potential induced by the three gates on the carbon atoms in CNT over the full extension of the device. Along with this, in the right part of Fig. 8, we show the corresponding transmission spectrum $T(E)$, for four gate voltages $V_{\mathrm{Gate-B}} = -2.0$ V, 1.0 V, 2.0 V, and 4.0 V, which represent significantly different locations on the conduction curves in Fig. 7.

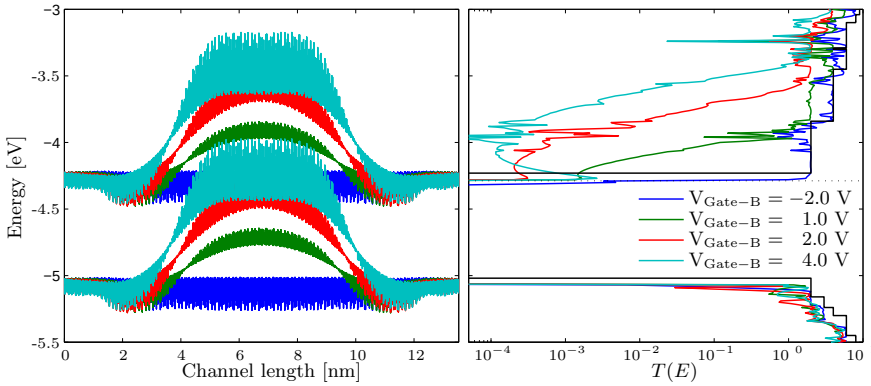


FIG. 8: (Color online) Fields induced along the length of the device (left panel) and the transmission spectrum (right panel) for the various gate voltages $V_{\mathrm{Gate-B}}$ from -2.0 V to 4.0 V as indicated.

From Fig. 8 we can see how the bands are shifted upwards by an increasing amount as the Gate-B voltage in turned up. To begin with, e.g., for $V_{\mathrm{Gate-B}} = 1$ V, this results in lower conduction since the conduction band bends away from the Fermi level, which is indicated by the dotted line. When the gate voltage is at $V_{\mathrm{Gate-B}} = 2$ V the valence band almost reaches the conduction band in which case BTB tunneling becomes possible. By increasing the gate voltage further more bands become available for BTB tunneling and the effect is visible as a steady increase in the calculated transmission $T(E)$ just above the Fermi level.

We would like to point out that the results presented here have all been calculated with the modified WFM method using parameters $\lambda_{\min} = 0.1$ and $l = 1$. Moreover, we have verified the transmission results presented in Fig. 8 by applying also the standard WFM method, and obtain identical transmissions curves to within at least three significant digits. We also emphasize, that the total CPU time required for calculating the 16 transmission spectra used for Fig. 7 (~ 53 hours) is less than half the time needed for calculating the four transmission curves with the standard WFM method (~ 141 hours) which were used for verification. The overall time saving achieved with the developed WFM method was therefore more than an order of magnitude for this application.

V. SUMMARY

We have developed an efficient approach for calculating quantum transport in nano-scale systems based on the WFM scheme originally proposed by Ando in reference [16]. In the standard implementation of the WFM method for two-probe systems, all bulk modes of the electrodes are required in order to represent the transmitted

and reflected waves in a complete basis. By extending the central region of two-probe system with extra electrode principal layers, we are able to exclude the vast majority of the evanescent bulk modes from the calculation altogether. Our final algorithm is therefore highly efficient, and most importantly, errors and accuracy can be closely monitored.

We have applied the developed WFM algorithm to a CNFET in order to study the mechanisms of band-to-band tunneling. The setup was inspired by reference [25] and the results of this paper confirmed. By measuring the CPU-times for calculating transmission spectra of the CNFET two-probe system and comparing to cost of the standard WFM method we have observed a speed-up by more than a factor of 10. We therefore believe that this is an ideal method to be used with ab initio transport schemes for large-scale simulations.

Acknowledgments

This work was supported by the Danish Council for Strategic Research (NABIIT) under grant number 2106–04–0017, "Parallel Algorithms for Computational Nano–Science".

* Electronic address: hhs@imm.dtu.dk

[1] S. Datta, *Quantum Transport: Atom to Transistor* (Cambridge University Press, Cambridge, UK, 2005).
[2] M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, Phys. Rev. B **65**, 165401 (2002).
[3] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, Phys. Rev. B **31**, 6207 (1985).
[4] Y. Meir and N. S. Wingreen, Phys. Rev. Lett. **68**, 2512 (1992).
[5] M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour, Science **278**, 252 (1997).
[6] S. V. Faleev, F. Léonard, D. A. Stewart, and M. van Schilfgaarde, Phys. Rev. B **71**, 195422 (2005).
[7] P. Pomorski, C. Roland, and H. Guo, Phys. Rev. B **70**, 115408 (2004).
[8] H. S. Gokturk, in *Nanotechnology, 2005. 5th IEEE Conference on* (2005), vol. 2, pp. 677–680.
[9] M. Stilling, K. Stokbro, and K. Flensberg, in *NSTI Nanotech 2006 Technical Proceedings* (2006), vol. 3, p. 39.
[10] A. Nitzan and M. A. Ratner, Science **300**, 1384 (2003).
[11] M. Di Ventra, S. T. Pantelides, and N. D. Lang, Phys. Rev. Lett. **84**, 979 (2000).
[12] K. Stokbro, J.-L. Mozos, P. Ordejon, M. Brandbyge, and J. Taylor, Comp. Mat. Sci. **27**, 151 (2003).
[13] N. D. Lang and P. Avouris, Phys. Rev. Lett. **84**, 358 (2000).
[14] B. Larade, J. Taylor, H. Mehrez, and H. Guo, Phys. Rev. B **64**, 075420 (2001).
[15] P. A. Khomyakov and G. Brocks, Phys. Rev. B **70**, 195402 (2004).
[16] T. Ando, Phys. Rev. B **44**, 8017 (1991).
[17] P. A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, and P. J. Kelly, Phys. Rev. B **72**, 035450 (pages 13) (2005).
[18] G. Brocks, V. M. Karpan, P. J. Kelly, P. A. Khomyakov, I. Marushchenko, A. Starikov, M. Talanana, I. Turek, K. Xia, P. X. Xu, et al., Ψ_k-Newsletter **80**, 144 (2007), URL http://www.psi-k.org/newsletters/News_80/newsletter_80.pdf.
[19] H. H. B. S. rensen, P. C. Hansen, D. E. Petersen, S. Skelboe, and K. Stokbro, Physical Review B (Condensed Matter and Materials Physics) **77**, 155301 (pages 12) (2008), URL http://link.aps.org/abstract/PRB/v77/e155301.
[20] K. Xia, M. Zwierzycki, M. Talanana, P. J. Kelly, and G. E. W. Bauer, Phys. Rev. B **73**, 064420 (pages 21) (2006).
[21] N. W. Ashcroft and D. N. Mermin, *Solid State Physics* (Brooks Cole, 1976).
[22] F. Guinea, C. Tejedor, F. Flores, and E. Louis, Phys. Rev. B **28**, 4397 (1983).
[23] M. P. Lopez Sancho, J. M. Lopez Sancho, J. M. L. Sancho, and J. Rubio, J. Phys. F. **15**, 851 (1985).
[24] P. S. Krstić, X.-G. Zhang, and W. H. Butler, Phys. Rev. B **66**, 205319 (2002).
[25] J. Appenzeller, Y.-M. Lin, J. Knoch, and P. Avouris, Phys. Rev. Lett. **93**, 196805 (2004).
[26] R. Hoffmann, The Journal of Chemical Physics **39**, 1397 (1963), URL http://link.aip.org/link/?JCP/39/1397/1.
[27] K. Stokbro and *et al.*, unpublished.
[28] D. S. Fisher and P. A. Lee, Phys. Rev. B **23**, 6851 (1981).
[29] Bloch's theorem[21] $\boldsymbol{\psi}_i = \lambda_k \boldsymbol{\psi}_{i-1}$ for the ideal electrodes defines the phase factors $\lambda_k \equiv e^{\imath q_k d}$, where q_k is the complex wave number and d is the layer thickness, which are referred to as Bloch factors throughout this paper.
[30] When using the Landauer formula in Eq. (1) it is assumed that the electrode Bloch states carry unit current in the conduction direction. This can be conveniently accommodated by flux-normalizing the Bloch states, i.e., $\boldsymbol{\phi}^{\pm}_{L,k} \to (d_L / v^{\pm}_{L,k})^{\frac{1}{2}} \boldsymbol{\phi}^{\pm}_{L,k}$, in the case of the left electrode.[28]
[31] We should point out that the metallic electrodes in the two-probe systems considered in Table I can be fully described by much smaller unit cells than indicated (often only a few atoms are needed) and therefore the time spend on computing the bulk states can be vastly reduced in these specific cases. For a general method, however, which supports CNTs, nano wires, etc. as electrodes, the timings are appropriate for showing the overall trend in the computational costs.

PAPER III

Krylov subspace method for evaluating the self-energy matrices in electron transport calculations

Hans Henrik B. Sørensen* and Per Christian Hansen
Department of Informatics and Mathematical Modelling, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark

Dan Erik Petersen and Stig Skelboe
Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark

Kurt Stokbro
Nano-Science Center, University of Copenhagen, Universitetsparken 5, Building D, DK-2100 Copenhagen, Denmark
(Received 29 August 2007; revised manuscript received 4 March 2008; published 1 April 2008; corrected 3 April 2008)

We present a Krylov subspace method for evaluating the self-energy matrices used in the Green's function formulation of electron transport in nanoscale devices. A procedure based on the Arnoldi method is employed to obtain solutions of the quadratic eigenvalue problem associated with the infinite layered systems of the electrodes. One complex and two real shift-and-invert transformations are adopted to select interior eigenpairs with complex eigenvalues on or in the vicinity of the unit circle that correspond to the propagating and evanescent modes of most influence in electron transport calculations. Numerical tests within a density functional theory framework are provided to validate the accuracy and robustness of the proposed method, which in most cases is an order of magnitude faster than conventional methods.

DOI: 10.1103/PhysRevB.77.155301 PACS number(s): 73.40.−c, 73.63.−b, 72.10.−d, 85.65.+h

I. INTRODUCTION

Quantum transport has been an important research subject for more than a decade due to the ever-growing interest in simulating and fabricating nanoscale electronic devices. In particular, the experimental and theoretical investigation of current-voltage (*I-V*) characteristics for molecules and atomic structures placed between conducting electrodes has attracted much effort.[1–11] Most theoretical approaches are based on the Landauer-Büttiker formulation of quantum transport,[12] where the electrical properties of a central interface are described by the transmission coefficients of a number of one-electron states propagating coherently through the system. The widely used Green's function method[13,14] and the wave function matching method[15–17] are two such techniques. To apply these in practice and determine the current through a device under finite bias, it is necessary to evaluate the bulk modes or, correspondingly, the self-energy matrices of each electrode for a considerable number of different energies (chemical potentials) and possibly k points.[18] In many cases, this represents the dominant part of the computational work associated with electron transport calculations, assuming that the Hamiltonian of the system has been provided.

In this paper we develop an efficient method for computing the self-energy matrices using an iterative Krylov subspace technique. The foundation of the method is the evaluation of the self-energy matrices for the semi-infinite electrodes from the solutions of the quadratic eigenvalue problem (QEP) that arises for infinite periodic systems. This approach has been suggested by Ando[19] and studied by several authors.[15,16,20–23] It has been shown[16,24] to be equivalent to well-established iterative and recursive schemes.[25,26] A disadvantage of the latter schemes from a computational point of view is the need to introduce a small imaginary part in the energy in order to ensure that the iterations converge to the correct retarded surface Green's function. This imaginary part forces complex arithmetic in the numerical algorithms used, which is not always the case in the eigenproblem approach.[15,19]

The key motivation for developing the proposed method is the physical observation that only the propagating and the slowly decaying evanescent modes in the bulk electrodes contribute to the transmission of electrons through a semiconductor device of some extension.[8] These modes correspond to the solutions of the QEP that have complex eigenvalues in the vicinity of the unit circle. As recently suggested by Khomyakov *et al.*,[15] this makes it plausible to generate reduced self-energy matrices on the basis of a few selected solutions of the QEP, which include all the electrode-device coupling information that is necessary to determine the correct transmission. To really exploit such an approach in practice, an algorithm to search for and compute *only* the desired quadratic eigenpairs is required.

We will here consider the Arnoldi method[27] combined with a shift-and-invert strategy in order to obtain the QEP solutions. These techniques have proven effective in obtaining selected interior eigenvalues of large-scale general complex eigenproblems.[28–30] In addition, the recent surge of papers studying the Arnoldi procedure applied specifically to polynomial matrix problems indicates that this is a successful technique to build the Krylov subspace for QEPs.[31–34] The algorithm we develop assumes real Hamiltonian matrices (generalization to the complex case is described in Appendix A 2), and targets the complex eigenvalues which are on or inside the unit circle by applying shift-and-invert spectral transformations to $\pm 1/\sqrt{2}$ and $\hat{\imath}/\sqrt{2}$, where $\hat{\imath}$ is the imaginary unit, and subsequently generating a Krylov subspace for each with the Arnoldi method. Ritz pairs obtained by projecting the QEP onto the three Krylov subspaces give good approximations to the eigenpairs with eigenvalues close to the corresponding shifts. We will show that this method of proceeding is both rigorous and efficient by applying it to various Hamiltonians obtained using density functional theory

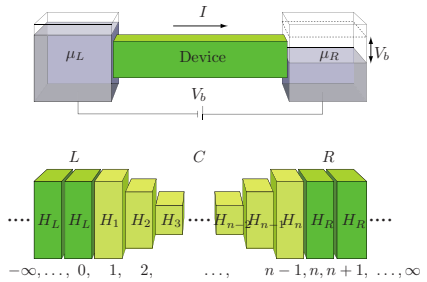


FIG. 1. (Color online) Schematic representation of a two-probe device with applied bias V_b. The top figure illustrates the Landauer-Büttiker picture of coherent scattering between electron reservoirs kept at chemical potentials μ_L and μ_R. The bottom figure shows the device part modeled by two semi-infinite electrodes (L and R) and a central region (C), each divided into principal layers that interact only with nearest-neighbor layers. The layers are described by square Hamiltonian matrices H_i of varying sizes and numbered $i=-\infty,\ldots,\infty$, as indicated.

(DFT) calculations with a localized basis of atomic orbitals.[35]

This paper is organized as follows. In Sec. II we give a brief exposition of our formalism for electron transport. The Krylov subspace method is introduced in Sec. III with details on its key parts: the Arnoldi method, the spectral transformations, and the convergence criterion. Typical convergence behavior is discussed in Sec. IV. The paper ends with numerical examples in Sec. V and a few concluding remarks.

II. ELECTRON TRANSMISSION AND SELF-ENERGY MATRICES

In this section we introduce our formalism, which combines the well-established Green's function method used for electron transport calculations[13,14,36] with the self-energy matrices obtained with the eigenvalue approach of Ando[19] as used in the wave function matching (WFM) method.[15–17] Our goal in combining the methods is to obtain, in the most efficient way, the spectrum of transmission coefficients $T(E)$ for two-probe systems (see top illustration in Fig. 1) in order to calculate the current $I=2e/h\int_{-\infty}^{\infty}T(E)[n_F(E-\mu_L)-n_F(E-\mu_R)]dE$ through the device, where E are the energies, n_F is the Fermi function, and μ_L and μ_R are the chemical potentials of the left (L) and right (R) electron reservoirs.[13,14]

A. Two-probe setup

Consider a two-probe system, as illustrated in the lower part of Fig. 1, where the device corresponds to the central region (C) and the reservoirs are two semi-infinite electrodes (L and R). The system has been divided into principal layers that interact only with nearest-neighbor layers and each layer is assumed to be described by appropriate Hamiltonian $\boldsymbol{H}_i$ and overlap $\boldsymbol{S}_i$ matrices, where i is the layer number, as represented, e.g., in a basis of localized nonorthogonal atomic orbitals. In this manner the Hamiltonian and overlap matrices are block-tridiagonal infinite matrices, where the off-diagonal blocks may be written $\boldsymbol{H}_{i,j}$ and $\boldsymbol{S}_{i,j}$. For the electrode Hamiltonian and overlap matrices we use subscripts L and R instead of numbers i,j. Notice also that the C region in this setup contains at least one layer of each electrode, which means that $\boldsymbol{H}_1=\boldsymbol{H}_L$ and $\boldsymbol{H}_n=\boldsymbol{H}_R$.

We refer the reader to Refs. 13, 14, and 36 for details on how to apply the Green's function method to the current setup. Here we limit ourselves to writing the primary results: First, the finite central region part of the infinite retarded Green's function matrix can be obtained as

$$\boldsymbol{G}_C^r=[(E+\hat{i}\eta)\boldsymbol{S}-\boldsymbol{H}_C-\boldsymbol{\Sigma}_L-\boldsymbol{\Sigma}_R]^{-1}, \tag{1}$$

where η is an infinitesimal quantity, $\boldsymbol{H}_C$ is the central region Hamiltonian, and the effect of the semi-infinite electrodes is accommodated through self-energy matrices $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_R$. Second, the total transmission coefficient $T(E)$ is then given by

$$T(E)=\mathrm{Tr}\{\boldsymbol{\Gamma}_L\boldsymbol{G}_C^r\boldsymbol{\Gamma}_R\boldsymbol{G}_C^a\}, \tag{2}$$

where $\boldsymbol{\Gamma}_{L/R}=\hat{i}(\boldsymbol{\Sigma}_{L/R}-\boldsymbol{\Sigma}_{L/R}^\dagger)$ are coupling matrices and $\boldsymbol{G}_C^a$ is the advanced central Green's function matrix, which is obtained from Eq. (1) by using $-\hat{i}\eta$ as the infinitesimal imaginary component in all terms (i.e., implicitly in $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_R$).

We find that an efficient approach (see Appendix A 1) to applying Eqs. (1) and (2) is to compute only a single *diagonal* block of $\boldsymbol{G}_C^r$ in order to evaluate $T(E)$. The question remains how to calculate the required self-energy matrices $\boldsymbol{\Sigma}_{L/R}$ in the most efficient manner.

B. Electrode self-energy matrices from QEPs

It is known that the surface Green's function matrices for a semi-infinite ideal electrode can be evaluated by recursive techniques that take 2^n-1 electrode layers into account in n iterations.[25,26] This is a fast and widely used approach to obtain the self-energy matrices when employing the Green's function method.[1,37]

Another approach has been proposed by Ando,[19] where one constructs and solves an appropriate QEP (introducing notation $\overline{\boldsymbol{H}}\equiv E\boldsymbol{S}-\boldsymbol{H}$)

$$\overline{\boldsymbol{H}}_{L,L}^\dagger\phi_k+\lambda_k\overline{\boldsymbol{H}}_L\phi_k+\lambda_k^2\overline{\boldsymbol{H}}_{L,L}\phi_k=0, \tag{3}$$

for $k=1,\ldots,2M_L$, where M_L is the number of orbitals local to the unit cell of the left electrode and similarly for the right electrode with $L\rightarrow R$. The procedure to determine the nontrivial solutions (i.e., the Bloch factors λ_k and electrode modes ϕ_k) from Eq. (3), and subsequently characterize these as propagating or evanescent, right-going (+) or left-going (−), is well described in the literature (we refer the reader to details in Refs. 15 and 16).

Applying Ando's approach via the formalism of the WFM method yields expressions[16]

$$\boldsymbol{\Sigma}_0^L=-\overline{\boldsymbol{H}}_{L,L}^\dagger(\boldsymbol{B}_L^-)^{-1}, \tag{4}$$

$$\boldsymbol{\Sigma}_{n+1}^R=-\overline{\boldsymbol{H}}_{R,R}\boldsymbol{B}_R^+ \tag{5}$$

for the electrode self-energy matrices in the layers 0 and n

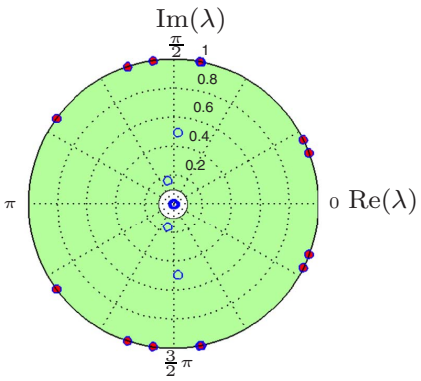


FIG. 2. (Color online) Positions of the 243 complex eigenvalues [blue (circles)] inside the unit disk for a Au(111) electrode with 27 atoms per unit cell at $E=-2$ eV. The 21 eigenvalues corresponding to propagating modes [red (filled) dots] are located on the unit circle. The modes of most significance in transmission calculations are located within the green (shaded) area given by $0.1 \leq |\lambda| \leq 1$.

$+1$ just *outside* the C region, where $\boldsymbol{B}_{L/R}^{\pm}$ are the Bloch matrices constructed from the solutions λ_k and ϕ_k [see the expressions in Ref. 16, in which the notation is $\mathbf{F}_{L/R}(\pm)$ for the Bloch matrices, and $\lambda_n(\pm)$ and $\boldsymbol{u}_n(\pm)$ for the solutions]. After evaluating these self-energy matrices we use them in the Green's function method described above (we set $\eta=0$ in this case, since the retarded Green's function is already uniquely defined by the self-energies[16,21]) and follow the steps outlined in Appendix A 1.

C. Reduced self-energy matrices

From a numerical perspective, it is convenient to keep only those eigenpairs from Eq. (3) that have eigenvalues λ_k within specific intervals[15]

$$\lambda_{\min} \leq |\lambda_k^+| \leq 1, \quad 1 \leq |\lambda_k^-| \leq \lambda_{\min}^{-1}, \tag{6}$$

for a reasonable choice of $\lambda_{\min}$. Evanescent modes with $|\lambda_k|$ outside these intervals are decaying or growing so fast that they have negligible influence in a two-probe setup like ours. The decisive factor in choosing $\lambda_{\min}$ is that the sets $\{\phi_k^+\}$ and $\{\phi_k^-\}$ of electrode modes included must be complete in the sense that they can fully represent the transmitted and reflected waves (cf. the WFM formalism).

In what follows, we exploit that a reasonable choice of $\lambda_{\min}$ for transmission calculations with our setup is often of the order 0.1.[38] For example, in the case of the polar plot in Fig. 2, where the Bloch factors with $|\lambda_k| \leq 1$ of a 27-atom Au(111) electrode unit cell are shown, the computationally significant modes can be identified as the eigenvalues inside the shaded area (i.e., by setting $\lambda_{\min}=0.1$). The numerical results given in Sec. V illustrate this observation quantitatively. A proper formal analysis is left for a future publication.[39]

III. KRYLOV SUBSPACE METHOD

In this section, we describe the Krylov subspace method for evaluating the electrode self-energy matrices $\boldsymbol{\Sigma}_0^L$ and $\boldsymbol{\Sigma}_{n+1}^R$. The crucial assumption in the approach is that we may strip the less important modes from the sets $\{\phi_k^+\}$ and $\{\phi_k^-\}$, and still obtain a good approximation to the self-energy matrix to be used in transmission calculations. For simplicity, we also assume that the electrode Hamiltonians are real, and give in Appendix A 2 a prescription to generalize to the complex case. Our current method, which targets the specific modes that are most important, can be characterized as a shift-and-invert Arnoldi method with adaptive subspace size. We will describe the key ingredients of the method: the Arnoldi procedure, the spectral transformations, and the convergence criterion. The goal is to present an alternative for obtaining the self-energy matrices, which is faster than existing techniques.

A. Arnoldi procedure

The Krylov subspace of dimension m generated by an $n \times n$ matrix $\boldsymbol{A}$ and an initial vector $\boldsymbol{v}_1$ is given by $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1) \equiv \text{span}\{\boldsymbol{v}_1, \boldsymbol{A}\boldsymbol{v}_1, \boldsymbol{A}^2\boldsymbol{v}_1, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{v}_1\}$.[40] In order to determine this space we apply the Arnoldi procedure[27] which generates an orthonormal basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ for $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1)$. We use the numerically most stable scheme that employs the modified Gram-Schmidt orthogonalization to successively construct the orthonormal vectors $\boldsymbol{v}_i$. Algorithm I below lists the steps of a continuable version of the Arnoldi procedure which is initially called with a parameter $k=1$ and a random starting vector $\boldsymbol{v}_1$. After $m-1$ iterations the $n \times m$ matrix $\boldsymbol{V}_m=(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ is available.

The projection of the matrix $\boldsymbol{A}$ onto $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1)$ is then $\boldsymbol{H}_m=\boldsymbol{V}_m^\dagger \boldsymbol{A} \boldsymbol{V}_m$, where $\boldsymbol{H}_m$ is $m \times m$ and upper Hessenberg (i.e., it has zeros below its lower bidiagonal). The matrix $\boldsymbol{H}_m$ is also constructed by Algorithm I. Approximate solutions of the eigenproblem $\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$ can subsequently be obtained as the so-called Ritz eigenpairs $(\gamma, \boldsymbol{V}_m\boldsymbol{y})$ of the projected eigenproblem $\boldsymbol{H}_m\boldsymbol{y}=\gamma\boldsymbol{y}$. As m increases the Ritz pairs become increasingly better approximations to certain eigenpairs of $\boldsymbol{A}$ (we point to Refs. 38 and 39 for details).

Algorithm I: Arnoldi procedure (continuable). Input: $k,m \in \mathbb{Z}$, $\boldsymbol{A} \in \mathbb{R}^{n,n}$, $\boldsymbol{V}_k \in \mathbb{R}^{n,k}$, $\boldsymbol{H}_k \in \mathbb{R}^{k,k}$. Output: $\boldsymbol{V}_{m+1} \in \mathbb{R}^{n,m+1}$, $\boldsymbol{H}_{m+1} \in \mathbb{R}^{m+1,m+1}$.

(1) If $k=1$, $\boldsymbol{v}_1=\boldsymbol{v}_1/\|\boldsymbol{v}_1\|_2$
(2) for $j=k,k+1,\ldots,m$ do
(3) $\boldsymbol{v}=\boldsymbol{A}\boldsymbol{v}_j$
(4) for $i=1,2,\ldots,j$ do
(5) $h_{ij}=\boldsymbol{v}_i^T\boldsymbol{v}$
(6) $\boldsymbol{v}=\boldsymbol{v}-h_{ij}\boldsymbol{v}_i$
(7) end
(8) $h_{j+1,j}=\|\boldsymbol{v}\|_2$
(9) if $h_{j+1,j}=0$, $m=j$, stop
(10) $\boldsymbol{v}_{j+1}=\boldsymbol{v}/h_{j+1,j}$
(11) end

One cannot know in advance how many steps will be needed before the eigenpairs of interest are well approximated by Ritz pairs. If many steps are necessary, then solving the projected eigenvalue problem becomes costly. More-

over, when applying our Krylov method to evaluate the self-energy matrices, we do not know the exact number of eigenpairs wanted and cannot estimate the required dimension of the Krylov subspace.

The first difficulty can be circumvented by restarting the Arnoldi method after a certain number of iterations using the obtained information to generate a better starting vector, or by deflating particular eigenvalues.[41] However, this will not improve on the second difficulty which requires an adaptive maximum dimension of the Krylov subspace. In addition, we observe in most of our applications that the gain from an efficient restart procedure (e.g., as devised by Morgan and Zeng[42]) does not outweigh the computational expense of the restarting overhead. The typical size of the self-energy matrices encountered is too small to make it beneficial to use such techniques, which have been developed for large-scale applications.

Therefore, we have chosen to employ a simple continuation scheme instead of restarting, where a check for convergence is performed after a given number of Arnoldi iterations, and if we are not satisfied, the procedure simply continues where it was left off. With the input parameter k, the listed Arnoldi algorithm is able to generate an initial Krylov subspace $\mathcal{K}_m$ of a given dimension m, but also to continue the process, augmenting the space with subsequent calls. This allows us to perform iterations as long as the approximations are unsatisfactory and/or there is doubt whether all wanted eigenpairs have been found.

An important special case to be considered when applying the Arnoldi procedure to solve an eigenvalue problem is that of algebraically multiple eigenvalues. A Krylov subspace method will, in theory, produce only one eigenvector corresponding to a multiple eigenvalue. So determination of multiplicity is quite difficult. Several approaches exist that deal with this problem, including deflation combined with effects of round-off error,[41] block Arnoldi procedures,[41] and so-called random restarts.[42,43] The present Krylov method does not incorporate any mechanisms to take algebraic multiplicity into account because such cases do not occur in practice for the applications of this work (eigenvalues will not be identical to machine precision in any of the numerical examples, but only to within $\sim 10-11$ digits; see Sec. IV).

B. Shift-and-invert transformations

Iterative methods based on Krylov subspaces produce Ritz values that converge fastest to the dominant part of the eigenvalue spectrum given by the extremal eigenvalues.[40] In the current application, it is the interior of the eigenvalue spectrum that is of interest, in particular the eigenvalues λ that satisfy $\lambda_{\min} \leq |\lambda| \leq \lambda_{\min}^{-1}$. To be able to find this part of the spectrum efficiently, we employ a shift-and-invert strategy which implies that the QEP in Eq. (3) is rewritten as

$$(\mu^2 \boldsymbol{M} + \mu \boldsymbol{C} + \boldsymbol{K})\boldsymbol{c}_0 = 0, \tag{7}$$

where

$$\boldsymbol{M} = \overline{\boldsymbol{H}}_{L,L}^T + \sigma \overline{\boldsymbol{H}}_L + \sigma^2 \overline{\boldsymbol{H}}_{L,L}, \tag{8}$$

$$\boldsymbol{C} = \overline{\boldsymbol{H}}_L + 2\sigma \overline{\boldsymbol{H}}_{L,L}, \tag{9}$$

$$\boldsymbol{K} = \overline{\boldsymbol{H}}_{L,L}, \tag{10}$$

and

$$\mu = \frac{1}{\lambda - \sigma}. \tag{11}$$

With this approach, the eigenvalues λ of Eq. (3) have been shifted by σ and inverted while the eigenvectors $\boldsymbol{c}_0$ are unchanged. Thus the dominant part of the spectrum of Eq. (7) now corresponds to the eigenvalues of the original QEP closest to the shift σ.

The simplest and currently state-of-the-art technique for solving Eq. (7) is by linearizing it to a generalized eigenvalue problem of twice the size.[44] In our case $\boldsymbol{M}$ is nonsingular and has size M_L. Therefore, a linearization results in a standard eigenvalue problem of size $2M_L$:

$$\boldsymbol{A}\boldsymbol{x} = \mu \boldsymbol{x}, \tag{12}$$

where $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} 0 & \boldsymbol{I} \\ -\boldsymbol{M}^{-1}\boldsymbol{K} & -\boldsymbol{M}^{-1}\boldsymbol{C} \end{pmatrix}, \tag{13}$$

and the $2M_L$ eigenvalues μ are identical to the ones of Eq. (7). The eigenvectors of Eq. (12) are given by $\boldsymbol{x}^T = (\boldsymbol{c}_0^T, \mu \boldsymbol{c}_0^T)$, so that the original eigenvectors $\boldsymbol{c}_0$ can be selected as the first M_L elements of $\boldsymbol{x}$.

If we assume that the Hamiltonian and overlap matrices for the electrodes are real, then the spectrum of the QEP in Eq. (3) is symmetric with respect to the real axis of the complex plane, and the eigenvalues either are real or occur in complex conjugate pairs.[44] In addition, as seen by transposing Eq. (3), the eigenvalues in this case also come in pairs, λ and $1/\lambda$. We will use these properties to present a simplified method for the extraordinary case of real $\overline{\boldsymbol{H}}_L$ and $\overline{\boldsymbol{H}}_{L,L}$, and subsequently discuss the steps required for the general complex case in Appendix A 2.

The purpose of the current method is thus to determine the eigenpairs $(\lambda, \boldsymbol{c}_0)$ of Eq. (3) that satisfy $\lambda_{\min} \leq |\lambda| \leq 1$ for a given $\lambda_{\min} > 0$ [the pairs that satisfy $1 \leq |\lambda| \leq \lambda_{\min}^{-1}$ can subsequently be obtained as $(\lambda^{-1}, \boldsymbol{c}_0)$]. As is apparent from the polar plot example in Fig. 2, the majority of the eigenvalues with $|\lambda| \leq 1$ are located near the origin. Therefore, it is not efficient to apply the shift $\sigma = 0$ in order to obtain the wanted eigenvalues, which lie in the outskirts of the unit disk. Instead we may apply four different shifts, given by $\sigma = \pm 1/\sqrt{2}$ and $\sigma = \pm \hat{i}/\sqrt{2}$, in four separate Arnoldi procedures. Each of these then covers a quarter slice of the unit disk and produces Ritz values that converge fast to eigenvalues close to the given shift. Simple sorting techniques can be employed in each Arnoldi procedure to take into account only the portion of the Ritz pairs that is covered by a given shift.

When applying the shift-and-invert strategy devised, two of the shifts have to be complex. In practice this means working in complex arithmetic or doubling the size of the problem.[45] However, in the case of real Hamiltonians it is advantageous to search for the complex eigenvalues in con-

jugate pairs and thereby eliminate one of the complex shifts. Moreover, this can be done almost entirely in real arithmetic as follows.

Notice that Eq. (12) was obtained by linearizing the shifted-and-inverted QEP written in Eq. (7). We may also reverse the order of the linearization and shift-and-invert operations. By performing, e.g., a first companion linearization of Eq. (3) that results in an eigenproblem $\hat{\boldsymbol{A}}\boldsymbol{x}=\lambda\boldsymbol{x}$ of double size, and subsequently a shift-and-invert transformation arriving at $(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}\boldsymbol{x}=\mu\boldsymbol{x}$, we see that the matrix applied in the Arnoldi procedures can also be written[44]

$$(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}=\begin{pmatrix}-\boldsymbol{M}^{-1}\hat{\boldsymbol{C}} & -\boldsymbol{M}^{-1}\boldsymbol{K}\\ \boldsymbol{I}-\sigma\boldsymbol{M}^{-1}\hat{\boldsymbol{C}} & -\sigma\boldsymbol{M}^{-1}\boldsymbol{K}\end{pmatrix}, \tag{14}$$

where

$$\hat{\boldsymbol{C}}=\overline{\boldsymbol{H}}_L+\sigma\overline{\boldsymbol{H}}_{L,L}. \tag{15}$$

The eigenpairs $(\mu,\boldsymbol{x})$ of $(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}\boldsymbol{x}=\mu\boldsymbol{x}$ are exactly the same as those of Eq. (12). In addition, we may now consider the combined spectral transformation for two conjugate shifts σ and σ^*, given by

$$\boldsymbol{T}=(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}(\hat{\boldsymbol{A}}-\sigma^*\boldsymbol{I})^{-1}=\frac{\mathrm{Im}\{(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}\}}{\mathrm{Im}\{\sigma\}}, \tag{16}$$

which was originally proposed by Parlett and Saad.[45] Applying the matrix $\boldsymbol{T}$ in the Arnoldi procedure generates approximate solutions to $\boldsymbol{T}\boldsymbol{x}=\mu'\boldsymbol{x}$, where the eigenvalues are given by

$$\mu'=\frac{1}{(\lambda-\sigma)(\lambda-\sigma^*)}, \tag{17}$$

which becomes extreme for conjugate eigenvalues λ and λ^* of Eq. (3) that are close to σ and σ^*. In our case, the complex shifts are purely imaginary: $\sigma=\hat{\imath}\beta$, where β is real. Then we have $\mu'=(\lambda^2+\beta^2)^{-1}$ and, more importantly, the matrix $\boldsymbol{T}$ is simply given by β^{-1} times the imaginary part of Eq. (14), written as

$$\boldsymbol{T}=\begin{pmatrix}-\beta^{-1}\,\mathrm{Im}\{\boldsymbol{M}^{-1}\hat{\boldsymbol{C}}\} & -\beta^{-1}\mathrm{Im}\{\boldsymbol{M}^{-1}\boldsymbol{K}\}\\ \mathrm{Re}\{\boldsymbol{M}^{-1}\hat{\boldsymbol{C}}\} & \mathrm{Re}\{\boldsymbol{M}^{-1}\boldsymbol{K}\}\end{pmatrix}, \tag{18}$$

which is purely real. This makes it feasible to use real arithmetic in all parts of the algorithm except for the initial complex LU factorization of $\boldsymbol{M}$, which is required for the matrix multiplications by $\boldsymbol{M}^{-1}$.

C. Algorithm and convergence criterion

The algorithm for our Krylov method is composed of two main parts, an iterative part that determines the wanted Ritz pairs $(\lambda,\boldsymbol{c}_0)$ which approximate the eigenpairs of the QEP in Eq. (3), and a noniterative part that sets up the Bloch matrices and evaluates the self-energy matrix from these by direct methods. The iterative part is organized as three independent computations, one for each of the used shifts σ. It consists of the application of the Arnoldi procedure together with a check for convergence plus the initial work to construct the input matrices for Algorithm I. As described in the previous section, the actual calculations will depend on whether the shift is real or imaginary.

The key steps of the Krylov method for evaluating the self-energy matrix $\boldsymbol{\Sigma}^L$ of the left electrode are presented in Algorithm II below. It is important to stress that the details of each step are kept at a minimum to enhance the readability. Furthermore, for evaluating the self-energy matrix $\boldsymbol{\Sigma}^R$ of the right electrode, the steps are exactly the same, except for the substitution $L\rightarrow R$ of all super- and subscripts and the removal of line 1 [this line is only required for left electrodes in order to obtain $\boldsymbol{\Sigma}^L$ from solutions $(\lambda^{-1},\boldsymbol{c}_0)$, e.g., by transposing Eq. (3)]. In the rest of this section we will discuss the main aspects of the algorithm.

Algorithm II: Krylov method to evaluate $\boldsymbol{\Sigma}^L$. Input: $m\in\mathbb{Z}$, $\lambda_{\min}\in[0,1]$, $\overline{\boldsymbol{H}}_L,\overline{\boldsymbol{H}}_{L,L},\overline{\boldsymbol{H}}_{L,L}^T\in\mathbb{R}^{M_L,M_L}$. Output: $\boldsymbol{\Sigma}^L\in\mathbb{C}^{M_L,M_L}$.

(1) Exchange matrices $\overline{\boldsymbol{H}}_{L,L}$ and $\overline{\boldsymbol{H}}_{L,L}^T$

(2) for $\sigma=1/\sqrt{2},-1/\sqrt{2},\hat{\imath}/\sqrt{2}$ do

(3) if σ is real, calculate $\boldsymbol{A}$ from Eq. (13)

else calculate $\boldsymbol{T}$ from Eq. (18) and set $\boldsymbol{A}=\boldsymbol{T}$

(4) select random vector $\boldsymbol{v}_1$ of size $2M_L$

(5) apply Algorithm I to generate $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1)$

(6) solve the projected eigenproblem $\boldsymbol{H}_m\boldsymbol{y}=\mu\boldsymbol{y}$

(7) if σ is real, select all $(\mu,\boldsymbol{y})$ that satisfy $\lambda_{\min}\leq|\mu^{-1}+\sigma|\leq1+\epsilon$, and store the Ritz pairs $(\lambda,\boldsymbol{c}_0)=(\mu^{-1}+\sigma,\boldsymbol{V}_m\boldsymbol{y})$ that have $\mathrm{Re}(\lambda)\mathrm{Re}(\sigma)\geq|\lambda|/2$

else select all $(\mu,\boldsymbol{y})$ that satisfy $\lambda_{\min}\leq|\mu^{-1}+\sigma^2|^{1/2}\leq1+\epsilon$, and evaluate the eigenvalues λ with the MR-2 method of Ref. 44 and store the Ritz pairs $(\lambda,\boldsymbol{c}_0)=(\lambda,\boldsymbol{V}_m\boldsymbol{y})$ that have $|\mathrm{Im}(\lambda)\mathrm{Im}(\sigma)|>|\lambda|/2$.

(8) for all stored Ritz pairs $(\lambda,\boldsymbol{c}_0)$, find residual $\|(\overline{\boldsymbol{H}}_{L,L}^T+\lambda\overline{\boldsymbol{H}}_L+\lambda^2\overline{\boldsymbol{H}}_{L,L})\boldsymbol{c}_0\|_2$, and check for convergence. If not satisfied, increase m appropriately and go to step 5

(9) end

(10) for all stored Ritz pairs $(\lambda,\boldsymbol{c}_0)$ having $(1+\epsilon)^{-1}\leq\lambda\leq1+\epsilon$, calculate group velocity v (see Ref. 15); discard the pairs with $v<0$ (i.e., the left-going modes)

(11) evaluate $\boldsymbol{B}_L^+$ and $\boldsymbol{\Sigma}^L=-\overline{\boldsymbol{H}}_{L,L}\boldsymbol{B}_L^+$ from the remaining pairs

First consider the steps 3–8 composing the body of the FOR loop, which are independently executed for the three given shifts σ. Each execution of these steps will determine Ritz pairs that are located in the corresponding quarter-slices of the unit disk. An illustration is shown in Fig. 3 for an Al(100) electrode, where the distinct slices are indicated by shaded areas and the current shifts by crosses. All wanted Ritz pairs found independently for the given shifts are assumed to be collected in a combined set when exiting the loop at step 9.

Initially, in step 3, the linearized and shifted-and-inverted matrix $\boldsymbol{A}$ to be applied in the Arnoldi procedure is determined from Eq. (13) if σ is real and from Eq. (18) if σ is complex. Then a starting vector $\boldsymbol{v}_1$ is selected randomly in step 4. A random starting vector is a reasonable choice in our case, where no prior information about the approximated eigenspace is available. In step 5 the Arnoldi procedure of Algorithm I is called to generate a Krylov subspace of size m, and in step 6, the corresponding eigenpairs $(\mu,\boldsymbol{y})$ of the

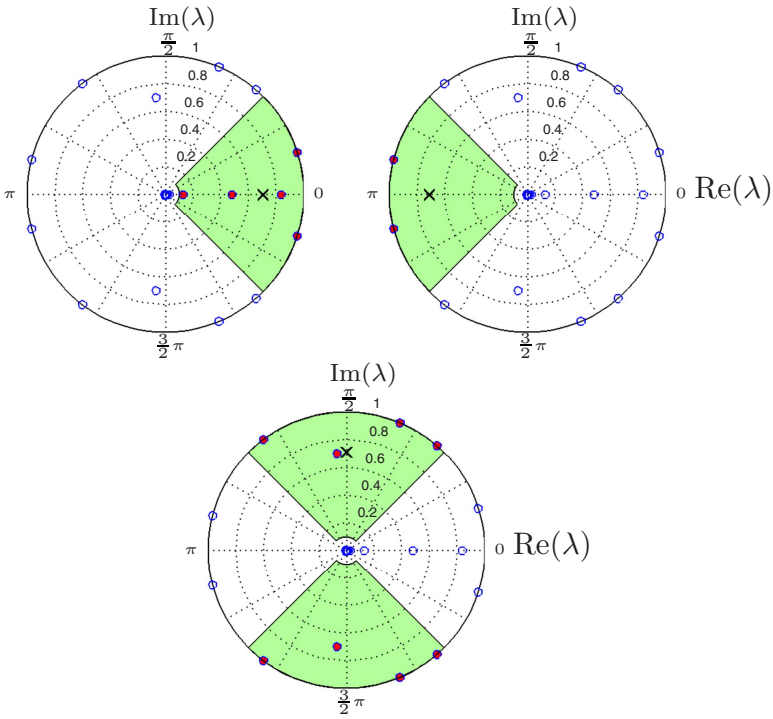


FIG. 3. (Color online) Illustration of the complex eigenvalues [blue (circles)] for the Al(100) electrode at $E=3$ eV. The eigenvalues corresponding to the wanted right-going modes [red (filled) dots] can be separated according to their location within three distinct green (shaded) areas of the unit disk and determined efficiently using shift-and-invert spectral transformations to $\pm 1/\sqrt{2}$ and $\hat{i}/\sqrt{2}$ (crosses).

shifted-and-inverted problem are found by solving the projected eigenproblem with a direct method. This is followed by an elaborate selection scheme to determine which of the available solutions $(\mu,\boldsymbol{y})$ correspond to wanted Ritz pairs $(\lambda,\boldsymbol{c}_0)$ that are located inside the valid quarter slice.

The selection scheme, as outlined in step 7, can be implemented as two separate processes. The first selection process is designed to identify those solutions $(\mu,\boldsymbol{y})$ that correspond to eigenpairs of the original QEP which satisfy $\lambda_{\min} \leq |\lambda| \leq 1$. It is important to realize, however, that, since all computations are done in finite-precision arithmetic, there is no guarantee that the propagating Bloch modes of the electrode will have magnitudes $|\lambda|$ exactly equal to 1. Even the left-going propagating modes that are targeted in our case can have $|\lambda|>1$. In practice, we therefore define the propagating modes to be those Ritz pairs $(\lambda,\boldsymbol{c}_0)$ that satisfy

$$(1+\epsilon)^{-1} \leq |\lambda| \leq 1+\epsilon \tag{19}$$

where ϵ is a small infinitesimal (set to 10^{-8} in our implementation). In order to make sure that all propagating modes are taken into consideration it is thus necessary to select all Ritz pairs that satisfy $\lambda_{\min} \leq |\lambda| \leq 1+\epsilon$.

To obtain the Ritz values λ used in the selection process, we have to transform the solutions $(\mu,\boldsymbol{y})$ of the projected eigenproblem to the corresponding Ritz pairs $(\lambda,\boldsymbol{c}_0)$ by reversing the shift-and-invert operation. The transformation again depends on whether the shift σ is real or imaginary. In the case of real σ, we have $\lambda=\mu^{-1}+\sigma$ from Eq. (11). For imaginary σ, Eq. (17) can be rearranged to $\lambda^2=\mu^{-1}+\sigma^2$, which has two solutions of equal magnitude. This is sufficient to allow selection on the basis of the magnitude $|\lambda|$; however, when it comes to obtaining the Ritz values λ themselves, it is necessary to use other means for imaginary σ, e.g., by forming the Rayleigh quotient.[40] In our case, and for QEPs in particular, it is possible and computationally advantageous to use alternatives to the Rayleigh quotient that work with vectors and matrices of size M_L instead of $2M_L$. Several such techniques that are both fast and accurate have recently been devised by Hochstenbach and van der Vorst.[46] We will adopt the MR-2 method of that paper, which yields $\lambda=\alpha/\beta$, for α and β defined as

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = -\widetilde{\boldsymbol{Z}}\overline{\boldsymbol{H}}_{L,L}^{T}\boldsymbol{c}_0, \tag{20}$$

where $\widetilde{\boldsymbol{Z}}$ is the pseudoinverse of $\boldsymbol{Z}=(\overline{\boldsymbol{H}}_{L,L}\boldsymbol{c}_0,\overline{\boldsymbol{H}}_{L}\boldsymbol{c}_0)$. Since all eigenvectors are unchanged by the shift-and-invert operation, the $\boldsymbol{c}_0$ vectors applied here are the first M_L elements of the Ritz vectors $\boldsymbol{V}_m\boldsymbol{y}$.

The remaining selection process in step 7 should single out the Ritz pairs that are inside the valid slice of the unit disk. To this end, we can apply the inner product of $(\mathrm{Re}\{\lambda\},\mathrm{Im}\{\lambda\})$ and $(\mathrm{Re}\{\sigma\},\mathrm{Im}\{\sigma\})$, given by

$$\mathrm{Re}\{\lambda\}\mathrm{Re}\{\sigma\}+\mathrm{Im}\{\lambda\}\mathrm{Im}\{\sigma\} = |\lambda||\sigma|\cos\,\theta, \tag{21}$$

where θ is the angle between λ and σ in a polar representation of the complex plane. In order for λ to be inside the

quarter slice that has σ on the bisector we must have $|\theta| \leq \pi/4$ or equivalently $\cos\theta \geq 1/\sqrt{2}$. For real shifts $\sigma = \pm 1/\sqrt{2}$, this observation yields the condition

$$\frac{\mathrm{Re}\{\lambda\}\mathrm{Re}\{\sigma\}}{|\lambda|} \geq \frac{1}{2}, \tag{22}$$

and similarly for imaginary shift $\sigma = \hat{\imath}/\sqrt{2}$,

$$\frac{|\mathrm{Im}\{\lambda\}\mathrm{Im}\{\sigma\}|}{|\lambda|} > \frac{1}{2}, \tag{23}$$

where the absolute value of the left-hand side is taken to allow λ to be in both the top and the bottom quarter slices. Notice that the equality is removed since the (very rare) event of λ lying exactly on the border of two slices is already taken into account in the condition for real σ.

In step 8 of Algorithm II the check for convergence is carried out. For each shift, the convergence condition is regarded as satisfied when all the Ritz pairs of interest that are also located inside the valid quarter slice are identified and accurate to a given tolerance. We estimate the accuracy of the obtained pairs $(\lambda, \boldsymbol{c}_0)$ by evaluating the corresponding relative residual norm, which yields the following convergence criterion:

$$\frac{\|(\overline{\boldsymbol{H}}_{L,L}^T + \lambda\overline{\boldsymbol{H}}_L + \lambda^2\overline{\boldsymbol{H}}_{L,L})\boldsymbol{c}_0\|_2}{\mathrm{norm}(\overline{\boldsymbol{H}}_L)} \leq \mathrm{tol} \tag{24}$$

where tol is the convergence tolerance and $\mathrm{norm}(\overline{\boldsymbol{H}}_L)$ is an appropriate norm for matrix $\overline{\boldsymbol{H}}_L$. In our implementation we set $\mathrm{tol}=10^{-11}$ and apply the approximation $\mathrm{norm}(\overline{\boldsymbol{H}}_L) \approx \|\mathrm{diag}(\overline{\boldsymbol{H}}_L)\|_2$, that is, we include only the diagonal entries of the two-norm of $\overline{\boldsymbol{H}}_L$. These choices require very low computational effort and give the correct result for all numerical examples we have investigated.

In the event that the convergence check in step 8 of Algorithm II is not satisfied, we assume that the dimension m of the Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v}_1)$ generated in step 5, is insufficient. Therefore, we increase m by some fixed amount and go back to step 5 to continue the Arnoldi procedure where it was left off. In the current implementation, we chose to increase the size of the Krylov subspace by $\Delta m = m/2$, where m is the initial value of m given as input. Our experiments show that, for optimal efficiency with this Δm, it is favorable to have the initial m within the range 30–50 if the sizes of the input matrices are of order less than 1000. After convergence has been achieved, the final steps 10–11 of Algorithm II present the operations required to collect the Ritz pairs that have been determined and subsequently obtain the self-energy matrix.

IV. TYPICAL CONVERGENCE BEHAVIOR

In this section, we briefly exemplify the typical convergence behavior of Algorithm II by monitoring the relative residual norm of the wanted eigenpairs as a function of the number of iterations. An expression for this norm for a given eigenpair $(\lambda, \boldsymbol{c}_0)$ is available as the left-hand side of Eq. (24). We will consider the Al(100) electrode at $E=3$ eV and parameter $\lambda_{\min}=0.1$, which requires a total of 13 eigenpairs to be determined (eight propagating modes and five evanescent modes) from the three separate Arnoldi procedures. This example corresponds to the situation illustrated in Fig. 3 and represents a typical calculation for an Al(100) electrode with 18 atoms per unit cell (the size of the self-energy matrix is 72).

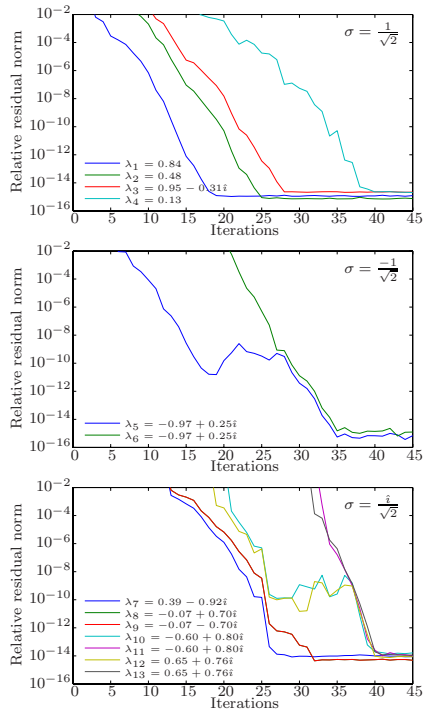


FIG. 4. (Color online) Convergence behavior of the Krylov algorithm for the Al(100) electrode at $E=3$ eV. The figures show the residual norm as a function of iterations for Ritz pairs that satisfy $0.1 \leq |\lambda| \leq 1+\epsilon$, in the case of shift-and-invert transformations to $\pm 1/\sqrt{2}$ and $\hat{\imath}/\sqrt{2}$, respectively.

In Fig. 4 we present curves showing the history of the residual norms for the wanted eigenpairs in each of the separate shift-and-invert Arnoldi procedures. We show only the 45 first iterations since this number is enough for convergence in all cases. Also, only residuals for eigenpairs corresponding to right-going modes are displayed.

The top figure of Fig. 4 illustrates the results from applying the shift $\sigma = 1/\sqrt{2}$ and shows that the Arnoldi procedure determines four different Ritz pairs with individual convergence curves. Comparing with the corresponding polar plot in Fig. 3 (top left), we observe a fifth eigenvalue $(\lambda = 0.95 + 0.31\hat{\imath})$ located inside the valid quarter slice. This fifth eigenvalue represents a left-going mode and is thus discarded in step 10 of Algorithm II. We also see by comparison with Fig. 3 that the eigenpair with eigenvalues furthest from the current shift (the cross) in the complex plane, in this case λ_4, is the slowest to converge.

The middle figure of Fig. 4 shows the convergence of the two Ritz pairs that are covered by the Arnoldi procedure with $\sigma=-1/\sqrt{2}$ and correspond to right-going modes in the present example. We note that λ_5 and λ_6 are nearly multiple eigenvalues, and that the behavior of the residual norms, where one eigenpair is available many iterations before its counterpart, is typical in such a case. Here, in particular, we see that eigenvalue λ_5 is determined to an accuracy of $\sim 10^{-11}$ after 18 iterations before λ_6 even shows up as a Ritz value of the projected eigenproblem. This indicates that λ_5 and λ_6 must be identical to around ten significant digits, and that they cannot be distinguished in our Arnoldi procedure before this accuracy is achieved. Without additional mechanisms to deal with multiple eigenvalues this then implies an upper bound condition on the value of the tol parameter.

The bottom figure of Fig. 4 shows the residual norm history of the remaining seven Ritz pairs required in the current example. These are determined by the Arnoldi procedure with imaginary shift $\sigma=\hat{i}/\sqrt{2}$ and correspond to filled dots in the bottom polar plot of Fig. 3 which represent right-going modes. We observe that the eigenvalue closest to σ, here denoted by λ_8, constitutes a complex conjugate pair together with λ_9, and that these have exactly the same residual norm curve (indistinguishable in the figure), although they are obtained separately as individual Ritz pairs in the algorithm.

In all residual norm figures, we see the trend that the eigenvalues located far from the position of the shift are slow to converge. This suggests that eigenvalues located in the vicinity of the intersections between the unit circle and the dividing lines of the four quarter slices will be the most difficult to determine since they are furthest from the corresponding shifts. The maximum distance from such an eigenvalue to σ is $1/\sqrt{2}$, which is the same as from σ to the origin. This raises concern whether the many unwanted eigenvalues close to the origin can become dominant compared to the wanted border eigenvalues. Fortunately, this is not the case because the unwanted eigenvalues close to the origin are clustered and therefore easy to represent in the Krylov subspace with only a few iterations.[40] We observe this in practice, e.g., from the bottom figure of Fig. 4, where the Ritz pair corresponding to λ_{12}, which lies close to the worst-case position on the unit circle, initially converges only slightly slower than the Ritz pair for λ_8 positioned right next to the shift.

V. NUMERICAL EXAMPLES

To illustrate the accuracy and practical aspects of the proposed Krylov subspace method we present transmission calculations for a metal-device-metal system that has been widely studied in the literature. In addition, we compute the current through this system at 1 and 2 V biases, and use the parameter $\lambda_{\min}$ to investigate the significance of the evanescent modes in obtaining the correct currents. Last, we apply the method to evaluate the self-energy matrices of a variety of electrodes (different types and sizes) and compare the actual measured CPU times[47] with those required by conventional methods.

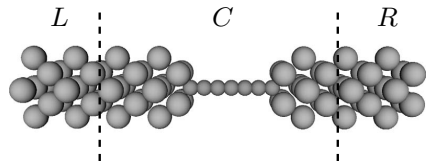


FIG. 5. Schematic illustration of the Al(100)-C7-Al(100) two-probe system.

A. Carbon wire between aluminum electrodes

To demonstrate the applicability of the proposed Krylov subspace method, we first consider carbon chains coupled to metallic electrodes, which have been investigated in detail recently.[1,5,6] Carbon atomic wires are interesting conductors since the equilibrium conductance of short monatomic chains varies with their length in an oscillatory fashion. We will examine the two-probe system shown in Fig. 5 corresponding to a straight wire of seven carbon atoms attached to Al(100) electrodes (lattice constant 4.05 Å). This structure exhibits a local maximum in the oscillatory conductance since it represents an odd-numbered C chain.[5] In our configuration, we fix the C-C distance to 2.5 bohrs and the distance between the ends of the carbon chain and the first plane of Al atoms at 1.0 Å. We use single-ζ basis sets for both types of atoms. The considered Al(100) electrode unit cell consists of 18 atoms in four layers with identical unit cells for the left and right electrodes. Notice that we do not use any symmetry properties of the metallic electrode to reduce the lateral size of the cells (as done, e.g., in Ref. 17) but rather use the full size matrices in Algorithm II. The same system has been studied by Brandbyge *et al.*[1]

We apply the proposed Krylov subspace method to calculate the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ of the left and right electrodes for a range of energies $E\in[-4\ \text{eV}, 4\ \text{eV}]$ and for different choices of the parameter $\lambda_{\min}$. The self-energy matrices are then used in the evaluation of the corresponding transmission coefficients $T(E)$.

Figure 6 presents the results for bias voltages $V_b=0$, 1, and 2 V in three cases of $\lambda_{\min}$. These significant bias settings are chosen for benchmarking and comparison reasons. The (black) full curves corresponding to $\lambda_{\min}=0.1$ reproduce the transmission spectra obtained in Ref. 1 (for 0 and 1 V) exactly except for the peak at $E=3.63$ eV (for 0 V), which is probably due to finer sampling in our work. In addition, we have calculated the similar curve with the full sets of electrode modes and the results are indistinguishable from those with the setting $\lambda_{\min}=0.1$ (and therefore not displayed in the figure). We note this as quantitative verification that the exclusion of the rapidly decaying evanescent modes is plausible in our setup.

We also see in Fig. 6 that the curves for the parameter $\lambda_{\min}$ set to 0.1 [black (full)] and 0.5 [red (dashed)] are almost identical, which indicates that the vast majority of the evanescent modes (those satisfying $|\lambda|<0.5$) have very little influence on $T(E)$ in the energy regime considered. However, when $\lambda_{\min}$ is set to 0.99 [blue (dotted curves)], in which case only propagating modes and very close to propagating modes are included in the evaluation of self-energy matrices,

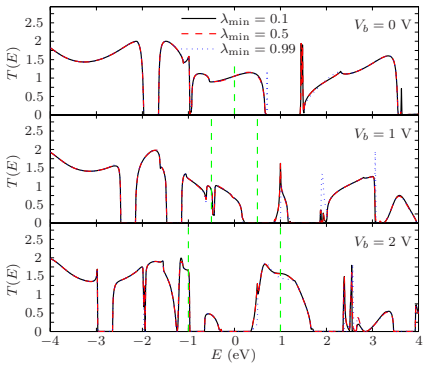


FIG. 6. (Color online) Transmission spectrum of the Al(100)-C7-Al(100) system for different bias voltages V_b. The self-energy matrices used in the $T(E)$ calculations have been obtained at the Γ point by the proposed Krylov subspace method with parameter $\lambda_{\min}$ at several settings: 0.1 [black (full) curve], 0.5 [red (dashed) curve], and 0.99 [blue (dotted) curve]. The bias windows are indicated by the vertical dashed lines.

there are several noticeable deviations from the other curves. Also inside the bias windows and especially for $V_b=2$ V, the disregard of the evanescent modes produces errors in the obtained transmission coefficients $T(E)$.

The deviations become even more evident in Fig. 7, where the current is displayed as a function of the parameter $\lambda_{\min}$ for nonzero bias voltages. As the value of $\lambda_{\min}$ is increased from around 0.5 to 1, the computed current I starts to

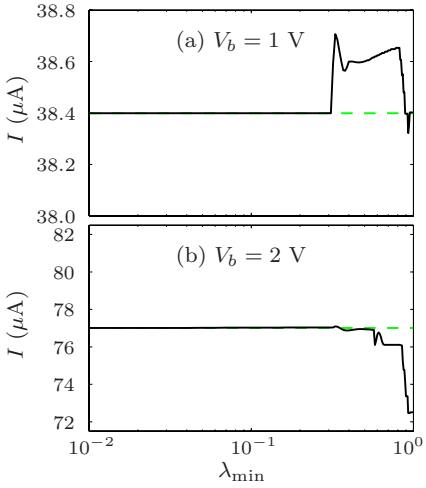


FIG. 7. (Color online) Current as a function of the parameter $\lambda_{\min}$ used by the Krylov subspace method for the Al(100)-C7-Al(100) system with applied bias voltages $V_b=$ (a) 1 and (b) 2 V. The correct currents obtained by conventional methods are $I \approx 38.4$ and ≈ 77.0 μA, respectively, indicated here by the green (dashed) lines.

TABLE I. CPU times in seconds for computing the left self-energy matrix $\mathbf{\Sigma}_L$ at 20 different energies E between −2 and 2 eV for selected electrode types and matrix sizes N. The parameter $\lambda_{\min}$ was set to 0.1.

Electrode type	Size	2^n iterative	DGEEV	Krylov
Li[a]	16	0.1	0.0	0.0
Fe[b]	54	4.2	2.3	0.6
Al(100)[c]	72	4.9	3.3	0.8
Al(100)[c]	128	27.9	17.5	3.6
Au(111)[d]	243	167.2	73.7	11.5
(2,2) CNT[e]	64	3.6	2.4	0.7
(4,4) CNT[e]	128	26.0	14.4	2.9
(8,8) CNT[e]	256	208.8	118.8	17.0
(12,12) CNT[e]	384	608.4	373.6	45.6
(16,16) CNT[e]	512	1230.0	1403.9	121.5
(20,20) CNT[e]	640	1542.3	1125.7	148.0

[a]Measurements from transmission calculations for ideal Li system.
[b]Measurements from transmission calculations for Fe-MgO-Fe; see geometry description in Ref. 10.
[c]Measurements from transmission calculations for Al(100)-C7-Al(100) described in this work (see also Ref. 1).
[d]Measurements from transmission calculations for Au(111)-BDT-Au(111); see, e.g., description in Ref. 11.
[e]Measurements from transmission calculations for ideal armchair (n,n) carbon nanotubes; see, e.g., description in Ref. 4.

depart significantly from the correct value. Therefore, we anticipate that at least some slowly decaying evanescent modes must be taken into account in order to describe the transmission properties of the Al(100)-C7-Al(100) system. Moreover, we see that this can be achieved in a rigorous and systematic fashion by selecting $\lambda_{\min}$ appropriately when using the proposed Krylov subspace method to calculate the self-energy matrices.

B. CPU run times

In this section we focus on the typical savings in the computational time that can be achieved when computing the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ with the proposed Krylov subspace method. We will compare run times directly with some conventional schemes usually applied in electron transport calculations. Our aim is to illustrate a significant speedup in calculating the self-energy matrices. This is of interest in future efforts to model much larger systems, and, in particular, for electrode unit cells that do not have any lateral symmetry properties.

Table I presents the profiling results when applying three different methods to calculate the same left self-energy matrix $\mathbf{\Sigma}_L$ for common types of electrodes and various matrix sizes N. In every case we consider only the Γ point and use single-ζ basis sets, except for Au(111) where a double-ζ-polarized set is used. Since the computational cost can vary significantly with E, the seconds listed represent the accumulated time of 20 independent calculations at equidistant energies in the interval $E \in [-2\ \text{eV}, 2\ \text{eV}]$. We focus on the

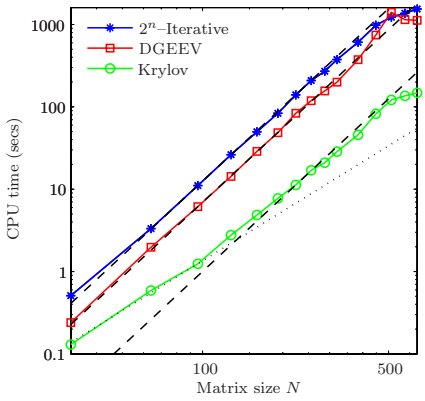


FIG. 8. (Color online) CPU times for computing the left self-energy matrix $\mathbf{\Sigma}_L$ plotted as a function of the size N of $\mathbf{\Sigma}_L$ for a range of armchair (n,n) CNT electrodes, where $n=1,\ldots,20$. The dotted and dashed lines indicate $O(N^2)$ and $O(N^3)$ computational complexity, respectively.

profiling for general electrode configurations and do not use lattice symmetries to reduce the order of the unit cells to elementary size even when this is possible.[17]

In the third column of Table I the run times to compute the correct self-energy matrices with the widely used iterative scheme of López Sancho *et al.*[26] are displayed. As the error in $\mathbf{\Sigma}_L$ obtained by this technique is reduced by $1/2^n$ after n iterations (we denote this method as 2^n iterative), it generally converges in $n\sim 22$ steps. In addition, run times for the conventional eigenvalue approach to evaluating the self-energy matrices, in which a standard eigensolver is used to determine the full set of modes, are presented in the fourth column. For this version, we simply substituted part of our Krylov subspace algorithm (steps 1–9 of Algorithm II) with the state-of-the-art LAPACK routine DGEEV.[47] In the last column the time required by the proposed Krylov subspace method is shown. In all cases of the latter the parameter $\lambda_{\min}$ was set to 0.1.

From the profiling results in Table I we see that the computational time of the Krylov subspace method is significantly reduced compared with the presently widely used 2^n-iterative technique. Also the conventional eigensolver scheme using DGEEV is typically faster than the 2^n-iterative algorithm [the exception for the (16,16) carbon nanotube (CNT) is related to cache usage[48]]. A comparison of the timings in the last two columns verifies that the cost to evaluate the self-energy matrices from only the few most important modes of the electrodes, as in our Krylov subspace method, is in general much lower than required by a direct eigensolver to determine all possible modes.

In order to illustrate the computational complexity of the methods we show the CNT run times as a function of the matrix size N in a logarithmic plot in Fig. 8. Clearly, all methods have $O(N^3)$ complexity; however, the Krylov subspace method initially follows the typical $O(N^2)$ complexity of the Arnoldi procedure[49] until the cost of the shift-and-invert operations becomes dominant. For $N>500$ we observe effects due to more and sometimes less favorable cache usage. Overall, we see that the Krylov subspace method is fastest by an order of magnitude for all but the smallest cases.

It is important to point out that the obtained self-energy matrices $\mathbf{\Sigma}_L$ are in all cases applied in a subsequent transmission calculation of $T(E)$ for the two-probe systems indicated in Table I, and the results then checked against those of the conventional methods [the resulting transmissions $T(E)$ are identical for the three methods in all cases of E to at least three decimals]. Furthermore, the setting of the parameter $\lambda_{\min}$ to 0.1 yields self-energy matrices evaluated from all the modes that have phases λ satisfying $0.1<|\lambda|<1+\epsilon$. This is more than adequate for obtaining correct results to an accuracy of three decimals for all the systems considered in this section. In practice, the parameter $\lambda_{\min}$ can often be selected >0.1 if lower accuracy in the $T(E)$ calculation is satisfactory, and this would show off the approach as even faster.

VI. CONCLUSIONS

In conclusion, we have developed an efficient and robust Krylov subspace method for evaluating the self-energy matrices that are required in electron transport calculations of nanoscale devices. The method exploits the observation that only the propagating and slowly decaying evanescent modes in the electrodes are computationally significant for determining the transmission coefficients when the system is appropriately set up.

The proposed method is based on the Arnoldi procedure and applies carefully chosen shift-and-invert spectral transformations to enhance the convergence toward the wanted interior eigenpairs that correspond to significant modes. We have investigated the convergence properties and shown that the accuracy and efficiency are mainly controlled by two parameters: the tolerance tol to be satisfied by of the relative residuals of the obtained Ritz values and the parameter $\lambda_{\min}$ that implicitly sets the number of modes taken into account.

In Sec. V we tested the Krylov subspace method on a metal-device-metal system and compared it to conventional methods. The applications show that the proposed method can be applied to calculate the transmission characteristics in a rigorous and systematic fashion and that the basic assumption of only including selective solutions in the electrode self-energy matrix is valid for many two-probe systems. The overall saving in computational time achieved by the Krylov subspace method is significant and in most cases more than an order of magnitude in comparison with conventional methods.

ACKNOWLEDGMENTS

The authors would like to thank J. Taylor and the people at Atomistix for helpful discussions. This work was supported by the Danish Council for Strategic Research (NABIIT) under Grant No. 2106-04-0017, "Parallel Algorithms for Computational Nano-Science."

APPENDIX A: COMPUTATIONAL DETAILS

1. Fast transmission calculation

We give the numerical steps to efficiently evaluate $T(E)$ via Eqs. (1) and (2). From the outset, the computational costs are reduced by taking into account that the self-energy matrices are nonzero only in the corner blocks, that is,

$$\boldsymbol{G}_C = \begin{pmatrix} \overline{\boldsymbol{H}}_1 - \boldsymbol{\Sigma}_1^L & \overline{\boldsymbol{H}}_{1,2} & & & \\ \overline{\boldsymbol{H}}_{1,2}^\dagger & \overline{\boldsymbol{H}}_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \overline{\boldsymbol{H}}_{n-1} & \overline{\boldsymbol{H}}_{n-1,n} \\ & & & \overline{\boldsymbol{H}}_{n-1,n}^\dagger & \overline{\boldsymbol{H}}_n - \boldsymbol{\Sigma}_n^R \end{pmatrix}^{-1}, \tag{A1}$$

where the self-energy blocks are numbered similarly to the Hamiltonian blocks. We then select a given diagonal block k and define self-energy matrices for every layer of the system, as[50–52]

$$\boldsymbol{\Sigma}_i^L = \overline{\boldsymbol{H}}_{i-1,i}^\dagger (\overline{\boldsymbol{H}}_{i-1} - \boldsymbol{\Sigma}_{i-1}^L)^{-1} \overline{\boldsymbol{H}}_{i-1,i}, \quad -\infty < i \le k, \tag{A2}$$

$$\boldsymbol{\Sigma}_i^R = \overline{\boldsymbol{H}}_{i,i+1} (\overline{\boldsymbol{H}}_{i+1} - \boldsymbol{\Sigma}_{i+1}^R)^{-1} \overline{\boldsymbol{H}}_{i,i+1}^\dagger, \quad k \le i < \infty, \tag{A3}$$

which can be used to recursively evaluate the self-energy matrices $\boldsymbol{\Sigma}_k^L$ and $\boldsymbol{\Sigma}_k^R$ when the matrices $\boldsymbol{\Sigma}_1^L$ and $\boldsymbol{\Sigma}_n^R$ (or $\boldsymbol{\Sigma}_0^L$ and $\boldsymbol{\Sigma}_{n+1}^R$ of the semi-infinite electrodes) are available. The kth block of the Green's function matrix is now given by

$$\boldsymbol{G}_{k,k} = (\overline{\boldsymbol{H}}_k - \boldsymbol{\Sigma}_k^L - \boldsymbol{\Sigma}_k^R)^{-1}, \tag{A4}$$

which corresponds to inverting the block of smallest size in the system, if k is chosen accordingly. Finally Eq. (2) is applied in a simplified version

$$T(E) = \mathrm{Tr}\{\boldsymbol{\Gamma}_k^L \boldsymbol{G}_{k,k} \boldsymbol{\Gamma}_k^R \boldsymbol{G}_{k,k}^\dagger\}, \tag{A5}$$

where the relation $\boldsymbol{G}_{k,k}^a = (\boldsymbol{G}_{k,k}^r)^\dagger$ between the advanced (a) and retarded (r) Green's functions is used [$\boldsymbol{G}^a = (\boldsymbol{G}^r)^\dagger$ is valid when E is real, since $\boldsymbol{H}$ is Hermitian and $\boldsymbol{\Sigma}^a = (\boldsymbol{\Sigma}^r)^\dagger$; see Ref. 13].

2. Generalization to complex Hamiltonian matrices and k-point sampling

In the Krylov subspace method presented in this paper we have assumed that the electrode Hamiltonian matrices are real in order to simplify the computational procedures. We now discuss the steps required to handle the case of complex $\overline{\boldsymbol{H}}_L$ and $\overline{\boldsymbol{H}}_{L,L}$, which is the case, e.g., when applying k-point sampling (Algorithm II works only for the Γ point).

As noted in Sec. III B, the assumption of real $\overline{\boldsymbol{H}}_L$ and $\overline{\boldsymbol{H}}_{L,L}$ leads to simplifications with the shift-and-invert operations:

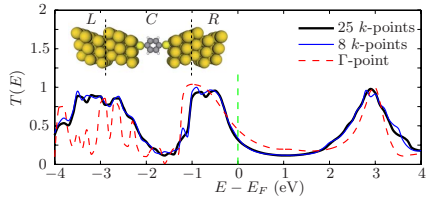


FIG. 9. (Color online) Transmission spectrum of the Au(111)-BDT-Au(111) system for different k-point samplings and V_b=0. The self-energy matrices used in the $T(E)$ calculations have been obtained by the generalized Krylov subspace method with parameter $\delta_{\min}$=0.1.

First, we may consider only right-going modes $(\lambda, \boldsymbol{c}_0)$ with $|\lambda| \le 1$ since the left-going modes are uniquely related as $(\lambda^{-1}, \boldsymbol{c}_0)$, and, second, we can use the spectral transformation $\boldsymbol{T}$ in Eq. (18) to determine the wanted eigenpairs for the two imaginary shifts $\sigma = \pm \hat{\imath}/\sqrt{2}$ simultaneously and in real arithmetic.

In order to generalize the Krylov subspace method to complex Hamiltonian matrices, it is thus necessary to determine the left-going modes satisfying $1 \le |\lambda| \le \lambda_{\min}^{-1}$ (i.e, located outside the unit circle) directly, since there is no general relation to the right-going modes (we note that it is advantageous to change the shift positions to be outside the unit circle, although this is not necessary for good convergence). Furthermore, we must abandon the $\boldsymbol{T}$ matrix and perform two independent shift-and-invert operations for $\sigma = \pm \hat{\imath}/\sqrt{2}$. It is clear that all this is now done in complex arithmetic and that the extra shift required will make the general algorithm a little more expensive (as shown in Sec. V B, the LU factorization required for each shift-and-invert operation is the dominant cost of our approach).

We have implemented the generalization and can illustrate its applicability by converging the transmission spectrum of the benzene di-thiol (BDT) molecule coupled to gold (111) surfaces in Fig. 9 by 3×3 and 7×7 k-point sampling of the Monkhorst type.[53] The calculation setup used is exactly the same as in Ref. 11 and the results can be confirmed.[3,11] Also, we have computed $T(E)$ for each E and k with self-energy matrices of both the 2^n-iterative method and the Krylov subspace method and checked that the results are identical to within three decimals. The CPU times required for, e.g., the 3×3 curve (eight k points) were 167 and 32 min for the two methods, respectively, while the Γ-point curve takes 2.7 min with Algorithm II. We conclude that the generalized Krylov subspace algorithm is, in this case, 1.5 times slower (per k point) than the real matrix version presented in Sec. III but still more than five times faster than the commonly used 2^n-iterative approach.

*hhs@imm.dtu.dk

[1] M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, Phys. Rev. B **65**, 165401 (2002).
[2] M. Di Ventra, S. T. Pantelides, and N. D. Lang, Phys. Rev. Lett. **84**, 979 (2000).
[3] S. V. Faleev, F. Léonard, D. A. Stewart, and M. van Schilfgaarde, Phys. Rev. B **71**, 195422 (2005).
[4] H. S. Gokturk, in *Proceedings of the Fifth IEEE Conference on Nanotechnology*, 2005, Vol. 2, pp. 677–680.
[5] N. D. Lang and P. Avouris, Phys. Rev. Lett. **84**, 358 (2000).
[6] B. Larade, J. Taylor, H. Mehrez, and H. Guo, Phys. Rev. B **64**, 075420 (2001).
[7] A. Nitzan and M. A. Ratner, Science **300**, 1384 (2003).
[8] P. Pomorski, C. Roland, and H. Guo, Phys. Rev. B **70**, 115408 (2004).
[9] M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour, Science **278**, 252 (1997).
[10] M. Stilling, K. Stokbro, and K. Flensberg, Mol. Simul. **33**, 557 (2007).
[11] K. Stokbro, J.-L. Mozos, P. Ordejon, M. Brandbyge, and J. Taylor, Comput. Mater. Sci. **27**, 151 (2003).
[12] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, Phys. Rev. B **31**, 6207 (1985).
[13] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge U.K., 1995).
[14] Y. Meir and N. S. Wingreen, Phys. Rev. Lett. **68**, 2512 (1992).
[15] P. A. Khomyakov and G. Brocks, Phys. Rev. B **70**, 195402 (2004).
[16] P. A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, and P. J. Kelly, Phys. Rev. B **72**, 035450 (2005).
[17] K. Xia, M. Zwierzycki, M. Talanana, P. J. Kelly, and G. E. W. Bauer, Phys. Rev. B **73**, 064420 (2006).
[18] K. S. Thygesen and K. W. Jacobsen, Phys. Rev. B **72**, 033401 (2005).
[19] T. Ando, Phys. Rev. B **44**, 8017 (1991).
[20] P. S. Krstić, X.-G. Zhang, and W. H. Butler, Phys. Rev. B **66**, 205319 (2002).
[21] D. H. Lee and J. D. Joannopoulos, Phys. Rev. B **23**, 4997 (1981).
[22] S. Sanvito, C. J. Lambert, J. H. Jefferson, and A. M. Bratkovsky, Phys. Rev. B **59**, 11936 (1999).
[23] T. Shimazaki, H. Maruyama, Y. Asai, and K. Yamashita, J. Chem. Phys. **123**, 164111 (2005).
[24] J. Velev and W. Butler, J. Phys.: Condens. Matter **16**, R637 (2004).
[25] F. Guinea, C. Tejedor, F. Flores, and E. Louis, Phys. Rev. B **28**, 4397 (1983).
[26] M. P. Lopez Sancho, J. M. Lopez Sancho, J. M. L. Sancho, and J. Rubio, J. Phys. F: Met. Phys. **15**, 851 (1985).
[27] W. E. Arnoldi, Q. Appl. Math. **9**, 17 (1951).
[28] M. N. Kooper, H. A. van der Vorst, S. Poedts, and J. P. Goedbloed, J. Comput. Phys. **118**, 320 (1995).
[29] K. Meerbergen and D. Roose, IMA J. Numer. Anal. **16**, 297 (1996).
[30] N. Nayar and J. M. Ortega, J. Comput. Phys. **108**, 8 (1993).
[31] Z. Bai and Y. Su, SIAM J. Matrix Anal. Appl. **26**, 640 (2005).
[32] L. Hoffnung, R.-C. Li, and Q. Ye, Linear Algebr. Appl. **415**, 52 (2006).
[33] U. B. Holz, G. H. Golub, and K. H. Law, SIAM J. Matrix Anal. Appl. **26**, 498 (2004).
[34] Q. Ye, Appl. Math. Comput. **172**, 818 (2006).
[35] First-principles DFT calculations are done with the commercial software package ATOMISTIX TOOLKIT 2.0. We use norm-conserved pseudopotentials for the core electrons and the local density approximation for the exchange-correlation potential (Ref. 1). More details about the software can be found on the company website (www.atomistix.com).
[36] P. N. C. Caroli, R. Combescot, and D. Saint-James, J. Phys. C **4**, 916 (1971).
[37] M. B. Nardelli, Phys. Rev. B **60**, 7828 (1999).
[38] A brief explanation for this is that, since the boundary layers of the *C* region in our setup are given by principal electrode layers, the evanescent modes that decay very fast do not "survive" the propagation through these layers and therefore do not give any components outside the sets $\{\phi_k^+\}$ and $\{\phi_k^-\}$ at the boundaries of *C*.
[39] H. H. B. Sørensen, D. E. Petersen, S. Skelboe, P. C. Hansen, and K. Stokbro (unpublished).
[40] L. N. Trefethen and D. Bau, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997).
[41] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide* (SIAM, Philadelphia, 2000).
[42] Ronald B. Morgan and M. Zeng, Linear Algebr. Appl. **415**, 96 (2006).
[43] Z. Jia, J. Comput. Math. **17**, 257 (1999).
[44] F. Tisseur and K. Meerbergen, SIAM Rev. **43**, 235 (2001).
[45] B. N. Parlett and Y. Saad, Linear Algebr. Appl. **88–89**, 575 (1987).
[46] M. E. Hochstenbach and H. A. van der Vorst, SIAM J. Sci. Comput. **25**, 591 (2003).
[47] All computations in this work were done on a Sun ULTRASPARC IV dual-core CPUs (1350 MHz/8 MB L2-cache). We use the vendor-supplied Sun Performance Library that includes platform-optimized versions of LAPACK routines.
[48] For the armchair (16,16) CNT electrode (N=512) the call to DGEEV produces an extremely high number of L2 cache misses, many more than for the larger (18,18) CNT electrode (N=576). This causes the very poor run times of the DGEEVmethod for this particular electrode.
[49] G. W. Stewart, *Matrix Algorithms* (SIAM, Philadelphia, 2001).
[50] E. M. Godfrin, J. Phys.: Condens. Matter **3**, 7843 (1991).
[51] D. E. Petersen, H. H. B. Sørensen, S. Skelboe, P. C. Hansen, and K. Stokbro, J. Comput. Phys. **227**, 3174 (2008).
[52] S. Y. Wu, J. Cocks, and C. S. Jayanthi, Phys. Rev. B **49**, 7957 (1994).
[53] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).